BUSINESS INTELLIGENCE

Técnicas, herramientas y aplicaciones

María Pérez Marqués

Alfaomega

BUSINESS INTELLIGENCE

Técnicas, herramientas y aplicaciones

María Pérez Marqués

Alfaomega

Diseño de colección, cubierta y pre-impresión: *GRUPO RC*

Datos catalográficos
Pérez, María
Business Intelligence. Técnicas, herramientas y aplicaciones
Primera Edición
Alfaomega Grupo Editor, S.A. de C.V., México
ISBN: 978-607-622-433-5
Formato: 17 x 23 cm Páginas: 432

Business Intelligence. Técnicas, herramientas y aplicaciones
María Pérez Marqués
ISBN: 978-84-943055-2-8 edición original publicada por RC Libros, Madrid, España
Derechos reservados © 2015 RC Libros

Primera edición: Alfaomega Grupo Editor, México, mayo 2015

© 2015 Alfaomega Grupo Editor, S.A. de C.V.
Pitágoras 1139, Col. Del Valle, 03100, México D.F.

Miembro de la Cámara Nacional de la Industria Editorial Mexicana
Registro No. 2317

Pág. Web: **http://www.alfaomega.com.mx**
E-mail: **atencionalcliente@alfaomega.com.mx**

ISBN: 978-607-622-433-5

Derechos reservados:
Esta obra es propiedad intelectual de su autor y los derechos de publicación en lengua española han sido legalmente transferidos al editor. Prohibida su reproducción parcial o total por cualquier medio sin permiso por escrito del propietario de los derechos del copyright.

Nota importante:
La información contenida en esta obra tiene un fin exclusivamente didáctico y, por lo tanto, no está previsto su aprovechamiento a nivel profesional o industrial. Las indicaciones técnicas y programas incluidos, han sido elaborados con gran cuidado por el autor y reproducidos bajo estrictas normas de control. ALFAOMEGA GRUPO EDITOR, S.A. de C.V. no será jurídicamente responsable por: errores u omisiones; daños y perjuicios que se pudieran atribuir al uso de la información comprendida en este libro, ni por la utilización indebida que pudiera dársele.

Edición autorizada para venta en México y todo el continente americano.

Impreso en México. Printed in Mexico.

Empresas del grupo:

México: Alfaomega Grupo Editor, S.A. de C.V. – Pitágoras 1139, Col. Del Valle, México, D.F. – C.P. 03100.
Tel.: (52-55) 5575-5022 – Fax: (52-55) 5575-2420 / 2490. Sin costo: 01-800-020-4396
E-mail: atencionalcliente@alfaomega.com.mx

Colombia: Alfaomega Colombiana S.A. – Calle 62 No. 20-46, Barrio San Luis, Bogotá, Colombia,
Tels.: (57-1) 746 0102 / 210 0415 – E-mail: cliente@alfaomega.com.co

Chile: Alfaomega Grupo Editor, S.A. – Av. Providencia 1443. Oficina 24, Santiago, Chile
Tel.: (56-2) 2235-4248 – Fax: (56-2) 2235-5786 – E-mail: agechile@alfaomega.cl

Argentina: Alfaomega Grupo Editor Argentino, S.A. – Paraguay 1307 P.B. Of. 11, C.P. 1057, Buenos Aires, Argentina, – Tel./Fax: (54-11) 4811-0887 y 4811 7183 – E-mail: ventas@alfaomegaeditor.com.ar

ÍNDICE

Introducción ..... **XI**

Capítulo 1. Inteligencia de negocios y sistemas de información. Informes ... **1**

Finalidad de los sistemas de información y origen del Business Intelligence... 1
Herramientas para la toma de decisiones en Business Intelligence.
OLAP, informes y minería 2
Almacenes de datos. Data Warehouse 4
Herramientas OLAP, ROLAP y MOLAP 10
Informes OLAP con CUBE y ROLLUP en SQL 15
Funciones de cubo en Excel 22

Capítulo 2. Microsoft y la inteligencia de negocios. Power BI. Informes con Power Query Y Power Pivot **25**

Introducción y componentes de Power BI 25
Usuarios de Power BI 27
Power Query 28
Power Pivot 44
Crear cálculos 55
Crear relaciones 58
Crear jerarquías 59

Capítulo 3. Informes con Power BI. Visualización con Power View y Power Map **63**

Power View 63
Informes sencillos 69
Informes detallados 75
Informes con segmentaciones de datos 79
Informes de comparaciones 82
Power Map 86

Capítulo 4. Business Intelligence con Microsoft BI SQL Server 91

Características de Business Intelligence en Microsoft BI SQL Server 91
PowerPivot para SharePoint 2013 92
PowerPivot para SharePoint 2010 94
Power View para SharePoint Server: crear, guardar e imprimir informes 96
Crear un informe en Power View para SharePoint Server 97
Abrir un informe existente de Power View 97
Guardar un informe 98
Permisos para Power View 98
Exportar a PowerPoint desde Power View en SharePoint 99
Actualizar los datos del informe 99
Imprimir vistas en un informe de Power View para SharePoint Server 100
Informes de Power View basados en modelos de datos 100
Descripción de los objetos de modelo multidimensional de Power View 101
Gráficos y otras visualizaciones en Power View 104
Gráficos circulares 106
Gráficos de dispersión y de burbujas 107
Gráficos de líneas, barras y columnas 108
Gráficos de barras 108
Gráficos de columnas 109
Gráficos de líneas 109
Mapas 109
Múltiplos: un conjunto de gráficos con los mismos ejes 110
Matrices 111
Tarjetas 111
Mosaicos 111
Analysis Services, Integration Services y Reporting Services 112
Informes de Reporting Services (SSRS) 113
SQL Server Integration Services 116
Analysis Services 118

Capítulo 5. Modelado predictivo con herramientas de IBM 123

Técnicas de modelado predictivo 123
Técnicas de modelado con IBM SPSS Modeler 129
Técnicas de modelado predictivo con IBM SPSS Modeler 130
El nodo Regresión Lineal: Modelo de regresión múltiple 131

 ©Alfaomega-RC Libros

El nodo Lineal: Modelo de regresión múltiple automático..... 136
El nodo Regresión Logística: Modelos de elección discreta 141
El nodo Regresión Logística: Regresión logística paso a paso..... 146
El nodo GenLin: Modelo Lineal General 149
El nodo Discriminante: Modelos de análisis discriminante 155

Capítulo 6. Modelado predictivo con herramientas de SAS..... 159

Técnicas de modelado predictivo con SAS Enterprise Miner 159
Modelo de regresión múltiple con el nodo Regression..... 160
El nodo Regression: Modelo Lineal General GLM 166
Modelos de elección discreta Logit y Probit con el nodo Regression 177

Capítulo 7. Modelado para la segmentación y perfilado con herramientas de IBM ... 181

Técnicas de modelado para la segmentación y obtención de perfiles..... 181
El análisis cluster como técnica de perfilado y segmentación..... 183
Clusters jerárquicos, secuenciales, aglomerativos y exclusivos (S.A.H.N.) ... 185
El dendograma en el análisis cluster jerárquico 186
Análisis cluster no jerárquico..... 186
Clustering mediante redes neuronales 189
Técnicas de segmentación con IBM SPSS Modeler 193
El nodo k-medias: Cluster no jerárquico..... 193
El nodo Bietápico: Cluster jerárquico 199
Análisis cluster con redes neuronales: Nodo Kohonen..... 204
Segmentación y perfilado con árboles de decisión 208
Características de los árboles de decisión..... 210
Tipos de árboles de decisión 212
Árboles CHAID 212
Árboles CART 213
Árboles QUEST..... 213
Árboles de decisión con IBM SPSS Modeler 214
El nodo C5.0..... 215
El nodo Árbol C&R 218
El nodo Quest 221
El nodo CHAID 224

Capítulo 8. Segmentación y perfilado con herramientas de SAS 229

Análisis cluster con SAS Enterprise Miner. El nodo Clustering 229
Análisis cluster con redes neuronales: nodo SOM/KOHONEN 236
El proceso de creación de árboles de decisión con SAS Enterprise Miner 244
Nodo Input Data Source para leer ficheros de datos y enlazarlos con SAS Enterprise Miner 245
Nodo Input Data Source y los tipos de variables 248
Enlace de nodos de un diagrama. El nodo Data Partition 250
Nodo de modelado para crear el árbol 252
Profundizando en el nodo TREE 256

Capítulo 9. Reglas de asociación con herramientas de IBM y SAS 267

Reglas de asociación 267
Reglas de asociación en IBM SPSS Modeler. El nodo A priori 269
Reglas de asociación en IBM SPSS Modeler. El nodo CARMA 276
Datos tabulares 279
Datos transaccionales 279
Nodo de asociaciones en SAS Enterprise Miner 283
Nodo de análisis de relaciones en SAS Enterprise Miner 290
Visor de resultados 294

Capítulo 10. Pronósticos y tendencias con herramientas de IBM 303

El algoritmo Serie Temporal de IBM SPSS Modeler 303
Componentes de una serie temporal 304
Valores atípicos 307
Funciones de autocorrelación y autocorrelación parcial 308
Transformaciones de series 309
Serie predictora 310
Nodo Modelos de series temporales 311
Resultados del modelo de serie temporal ajustado 326
Generación de modelos de series temporales 337

Capítulo 11. Pronósticos y tendencias con herramientas de SAS 341

Predicciones incondicionales deterministas y suavizado con el modelizador de SAS 341

 ©Alfaomega-RC Libros

Modelos ARIMA con el modelizador automático de SAS 355
Modelos ARIMA estacionales y generales con el modelizador automático de SAS 362

Capítulo 12. Modelos automáticos y evaluación de modelos con herramientas de IBM y SAS 367

Modelización automática en IBM SPSS Modeler 367
El nodo clasificador automático de IBM SPSS Modeler 368
El algoritmo autonumérico de IBM SPSS Modeler 379
El algoritmo autoconglomeración de IBM SPSS Modeler 388
Fase de evaluación en SAS Enterprise Miner. Nodo Assessment 398
Trabajo con el nodo de evaluación 398
El fichero de datos con las predicciones 398
Models 399
Options 402
Reports 402
Output 402
Gráficos del nodo Assessment 403
Lift (Gráfico de ganancias) 403
Vista de los datos relativos a los gráficos de ganancias 406
Diagnostic (Gráfico de clasificación) 406
Threshold-Based (Gráfico de clasificación basado en el umbral) 407
Curvas ROC 409
Scatter Plot (Gráfico de dispersión) 410
Ejemplo de combinación de modelos de árboles con otros modelos 411

Índice analítico 415

INTRODUCCIÓN

Podemos definir la Inteligencia de negocio o BI (*Business Intelligence*) como el conjunto de estrategias enfocadas a la administración y creación de conocimiento sobre el medio, a través del análisis de los datos existentes en una organización o empresa.

El término ***inteligencia de negocio*** se refiere al uso de datos en una empresa para facilitar la toma de decisiones. Abarca tanto la comprensión del funcionamiento actual de la empresa, como la anticipación de acontecimientos futuros, con el objetivo de ofrecer conocimientos para respaldar las decisiones empresariales. Estamos así ante el proceso de ***extracción del conocimiento*** o KDD.

Las herramientas de inteligencia de negocios se basan en la utilización de un sistema de información de inteligencia que se forma con distintos datos extraídos de los datos de producción, con información relacionada con la empresa o sus ámbitos y con datos económicos. Mediante las herramientas y técnicas ELT (extraer, cargar y transformar), o actualmente ETL (extraer, transformar y cargar) se extraen los datos de distintas fuentes, se depuran y preparan (homogeneización de los datos) para luego cargarlos en un almacén de datos. La vida o el periodo de éxito de un software de inteligencia de negocios dependerá únicamente del éxito de su uso en beneficio de la empresa. Si esta empresa es capaz de incrementar su nivel financiero, administrativo y sus decisiones mejoran la actuación de la empresa, el software de inteligencia de negocios seguirá presente mucho tiempo, en caso contrario será sustituido por otro que aporte mejores y más precisos resultados.

Asimismo, las herramientas de inteligencia analítica posibilitan el modelado de las representaciones basadas en consultas para crear un cuadro de mando integral que sirve de base para la presentación de informes.

El conjunto de herramientas y metodologías para la inteligencia de negocios tiene en común las siguientes características:

- *Accesibilidad a la información*. Los datos son la fuente principal de este concepto. Lo primero que deben garantizar este tipo de herramientas y técnicas será el acceso de los usuarios a los datos con independencia de su procedencia.

- *Apoyo en la toma de decisiones*. Se busca ir más allá en la presentación de la información, de manera que los usuarios tengan acceso a herramientas de análisis que les permitan seleccionar y manipular sólo aquellos datos que les interesen.

- *Orientación al usuario final*. Se busca independencia entre los conocimientos técnicos de los usuarios y su capacidad para utilizar estas herramientas.

Una clasificación muy simple de las soluciones de Business Intelligence podría ser la siguiente:

- Informes
  - Informes predefinidos
  - Informes a medida
  - Consultas (*Query*) / Cubos OLAP (*On-Line Analytic Processing*).
  - Alertas

- Análisis
  - Análisis estadístico
  - Pronósticos (*Forecasting*)
  - Modelado predictivo o Minería de datos (*Data Mining*)
  - Optimización

La estrategia de la inteligencia de negocios debe ser vista como un proceso creativo, buscar nuevas formas de hacer las cosas, de generar valor en el mundo de continuo cambio, y ser efectivo en el corto plazo, por lo cual se necesita:

 ©Alfaomega-RC Libros

- Inteligencia para crear y compartir el conocimiento.
- La habilidad para integrar y administrar este conocimiento.
- La imaginación para visualizar acciones alternativas a las usuales y analizar sus consecuencias.
- La pericia para manejar los recursos y atender las necesidades actuales sin dejar de construir el futuro deseable.

Con la globalización, la competencia se acentúa y hay que reaccionar con rapidez, sorpresa y anticipación. Las herramientas de inteligencia de negocios desempeñan un papel muy importante en este proceso de superación y competitividad en que está inmerso el mercado actual.

Este libro analiza las herramientas más habituales en inteligencia de negocios y sus posibilidades de trabajo. Se utilizará el software adecuado para apoyar las técnicas de Business Intelligence para una extracción óptima del conocimiento contenido en los datos.

A través de ejemplos totalmente resueltos a lo largo del libro se irán presentando las diferentes técnicas de inteligencia de negocios. Los archivos se encuentran en la página web: www.rclibros.es en la sección *Zona de archivos*.

Se trata de exponer con sencillez y mediante una metodología interactiva los conceptos de extracción del conocimiento e inteligencia de negocios.

©Alfaomega-RC Libros

CAPÍTULO 1

INTELIGENCIA DE NEGOCIOS Y SISTEMAS DE INFORMACIÓN. INFORMES

FINALIDAD DE LOS SISTEMAS DE INFORMACIÓN Y ORIGEN DEL BUSINESS INTELLIGENCE

La información ***reduce nuestra incertidumbre*** (sobre algún aspecto de la realidad) y, por tanto, nos permite tomar mejores decisiones.

Inicialmente la finalidad de los sistemas de información era recopilar información sobre una parcela del mundo para ayudar en la toma de decisiones y se basaba en recuentos, censos civiles y militares, libros contables, etc. Actualmente, con la informatización de las organizaciones y la aparición de aplicaciones software operacionales sobre el sistema de información, la finalidad principal de los sistemas de información es dar soporte a los procesos básicos de la organización (ventas, producción, personal, etc.).

Una vez satisfecha la necesidad de tener un soporte informático para los procesos básicos de la organización (**sistemas de información para la gestión**), las organizaciones exigen nuevas prestaciones de los sistemas de información (**sistemas de información para la toma de decisiones**). Es aquí donde aparece el Business Intelligence.

HERRAMIENTAS PARA LA TOMA DE DECISIONES EN BUSINESS INTELLIGENCE. OLAP, INFORMES Y MINERÍA

Ante el problema de la toma de decisiones han aparecido diferentes herramientas de inteligencia de negocio o DSS que *coexisten*: EIS, OLAP, consultas & informes, minería de datos, etc.

Un EIS (*Executive Information System*) es un sistema de información y un conjunto de herramientas asociadas que tiene las siguientes carcaterísticas:

- Proporciona a los directivos acceso a la información de estado y sus actividades de gestión.
- Está especializado en analizar el estado diario de la organización (mediante indicadores clave) para informar rápidamente sobre *cambios* a los directivos.
- La información solicitada suele ser, en gran medida, numérica (*ventas semanales, nivel de stocks, balances parciales, etc.*) y representada de forma gráfica al estilo de las hojas de cálculo.

Las *herramientas OLAP* (*On-Line Analyitical Processing*) son más genéricas:

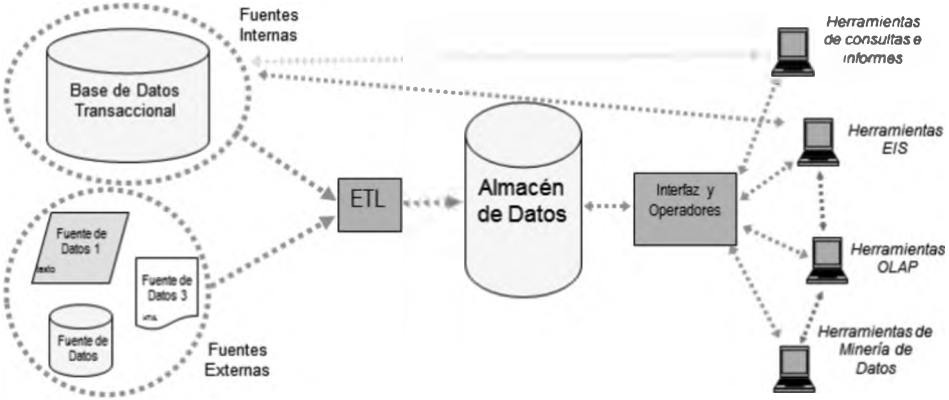
- Funcionan sobre un sistema de información (transaccional o almacén de datos).
- Permiten realizar agregaciones y combinaciones de los datos de maneras más complejas y ambiciosas, con objetivos de análisis más estratégicos.
- Están basadas, generalmente, en sistemas o *interfaces multidimensionales*.
- Se utilizan operadores específicos (además de los clásicos): *drill, roll, pivot, slice & dice,...*
- El resultado se presenta de una manera matricial o híbrida.
- Proporcionan facilidades para "manejar" y "transformar" los datos.
- **Producen otros "datos"** (más agregados, combinados).
- Ayudan a analizar los datos porque producen *diferentes vistas* de los mismos.

©Alfaomega-RC Libros

Por otro lado, los *sistemas de informes* o *consultas avanzadas* están basados, generalmente, en sistemas relacionales u objeto-relacionales, utilizan los operadores clásicos como concatenación, proyección, selección, agrupamiento... (en SQL y extensiones) y el resultado se presenta de una manera tabular.

Instrumentos más avanzados para la toma de decisiones en inteligencia de negocios son las herramientas de Minería de Datos. Son muy variadas y permiten "extraer" patrones, modelos, descubrir relaciones, regularidades, tendencias, etc. También producen "reglas" o "patrones" ("conocimiento").

La interrelación entre todas estas herramientas se presenta en el esquema siguiente:



Mediante las herramientas y técnicas ELT (extraer, cargar y transformar), o actualmente ETL (extraer, transformar y cargar) se extraen los datos de distintas fuentes externas e internas (bases de datos transaccionales), se depuran y preparan (homogeneización de los datos) para luego cargarlos en un almacén de datos. En el centro del esquema aparece el *almacén de datos*, que es el "**sistema de información central**" en todo este proceso. Un almacén de datos es una colección de datos orientada a un dominio, integrada, no volátil y variante en el tiempo para ayudar en la toma de decisiones. A partir del almacén de datos, mediante interfaces y operadores se utilizan las herramientas de informes, EIS, OLAP y Minería de Datos.

Los almacenes de datos y las técnicas OLAP son las maneras más efectivas y tecnológicamente más avanzadas para **integrar, transformar y combinar los datos para facilitar** al usuario o a otros sistemas **el análisis de la información**. La tecnología

OLAP generalmente se asocia a los almacenes de datos, aunque podemos tener almacenes de datos sin OLAP, y viceversa.

La minería de datos es solo una etapa del proceso de extracción de conocimiento a partir de datos. Consta de varias fases: Preparación de Datos (selección, limpieza y transformación), Análisis de Datos, Evaluación, Difusión y Uso de Modelos. Incorpora diferentes técnicas como árboles de decisión, regresión lineal, redes neuronales artificiales, técnicas predictivas, técnicas de segmentación, etc. Se aplica en campos diversos como el aprendizaje automático e inteligencia artificial, estadística, bases de datos, clasificación, categorización, estimación y regresión, agrupamiento, etc.

Los almacenes de datos no son *imprescindibles* para hacer extracción de conocimiento a partir de datos. Se puede hacer minería de datos sobre un simple fichero de datos. Las ventajas de organizar un almacén de datos para realizar minería de datos se amortizan sobradamente a medio y largo plazo cuando tenemos grandes volúmenes de datos, o estos aumentan con el tiempo, o provienen de fuentes heterogéneas o se van a combinar de maneras arbitrarias y no predefinidas.

ALMACENES DE DATOS. DATA WAREHOUSE

Generalmente, la información que se quiere investigar sobre un cierto dominio de la organización se encuentra en bases de datos y otras fuentes muy diversas, tanto internas como externas. Muchas de estas fuentes son las que se utilizan para el trabajo diario (*bases de datos operacionales*). Sobre estas mismas bases de datos de trabajo ya se puede extraer conocimiento (visión tradicional).

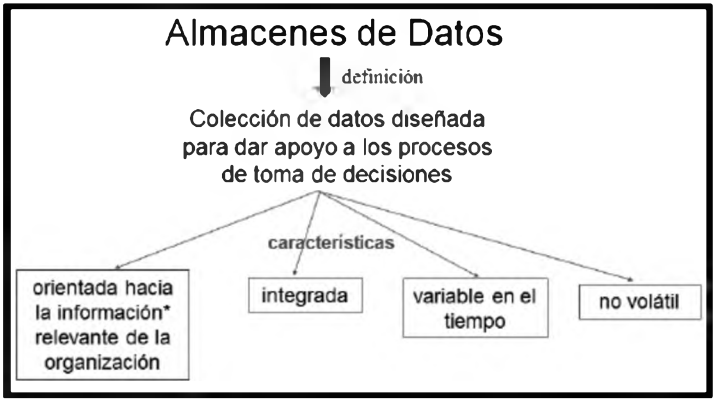
Una *base de datos transaccional* es una fuente de datos mediante la cual se mantiene el trabajo transaccional diario de los sistemas de información originales (conocido como OLTP, *On-Line Transactional Processing*). También se hacen análisis de los datos en tiempo real sobre la misma base de datos (conocido como OLAP, *On-Line Analytical Processing*).

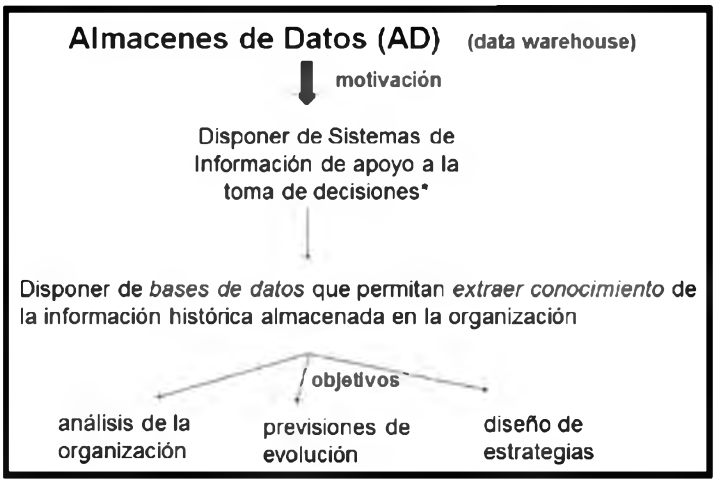
Como problemas más comunes, la base de datos transaccional perturba el trabajo transaccional diario de los sistemas de información originales ("killer queries"). Se debe hacer por la noche o en fines de semana. Además, la base de datos está diseñada para el trabajo transaccional, no para el análisis de los datos. Generalmente no puede ser en tiempo real.

©Alfaomega-RC Libros

Para operar eficientemente con los datos, los costes de almacenamiento masivo y conectividad se han reducido drásticamente en los últimos años. Parece razonable recopilar los datos (información histórica) en un sistema separado y específico. Aparece así el *Data warehouse* (Almacén o Bodega de Datos).

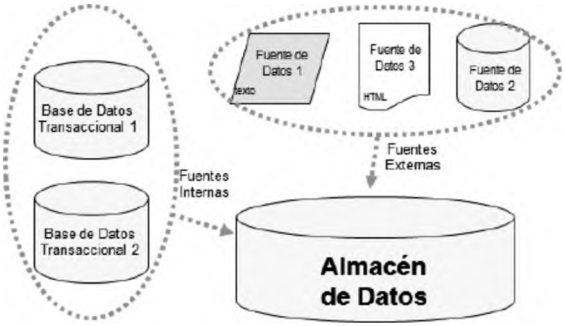
Los esquemas siguientes muestran la definición, la motivación y los objetivos de los almacenes de datos.



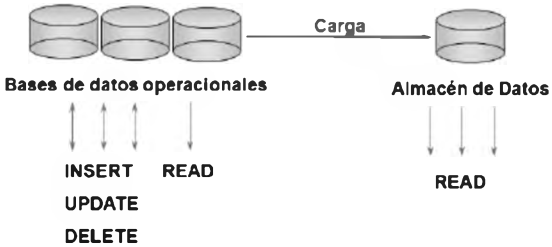


Un almacén de datos siempre está orientado hacia la información relevante de la organización. Se diseña para consultar eficientemente información relativa a las actividades (ventas, compras, producción...) básicas de la organización y no para soportar los procesos que se realizan en ella (gestión de pedidos, facturación, etc.).

Un almacén de datos integra datos recopilados de diferentes sistemas operacionales de la organización, incluyendo bases de datos transaccionales y/o fuentes externas.



Los datos en un almacén de datos son relativos a un periodo de tiempo y deben ser incrementados periódicamente. Los datos son almacenados como fotos (*snapshots*) correspondientes a periodos de tiempo. Además, los datos almacenados no son actualizados, solo son incrementados. Las operaciones de inserción, actualización y borrado de los datos se realizan en la base de datos operacional antes de que sean cargados en el almacén de datos.



Los almacenes de datos presentan múltiples ventajas para las organizaciones entre las que destacan la rentabilidad de las inversiones realizadas para su creación, el aumento de la competitividad en el mercado y el aumento de la productividad de los técnicos de dirección. Pero también presentan problemas como la infravaloración del esfuerzo necesario para su diseño y creación, la infravaloración de los recursos necesarios para la captura, la carga y el almacenamiento de los datos, el incremento continuo de los requisitos de los usuarios y la privacidad de los datos. El esquema siguiente presenta las diferencias esenciales entre una base de datos operacional y un almacén de datos.

©Alfaomega-RC Libros

Sistema Operacional (OLTP)	Almacén de datos (DW)
- almacena datos actuales	- almacena datos historicos
- almacena datos de detalle	- almacena datos de detalle y datos agregados a distintos niveles
-bases de datos medianas (100Mb-1Gb)	- bases de datos grandes (100Gb-1Tb)
- los datos son dinamicos (actualizables)	- los datos son estaticos
- los procesos (transacciones) son repetitivos	- los procesos no son previsibles
- el numero de transacciones es elevado	- el numero de transacciones es bajo o medio
- tiempo de respuesta pequeño (segundos)	- tiempo de respuesta variable (segundos-horas)
- dedicado al procesamiento de transacciones	- dedicado al análisis de datos
- orientado a los procesos de la organización	- orientado a la información relevante
- soporta decisiones diarias	- soporta decisiones estratégicas
- sirve a muchos usuarios (administrativos)	- sirve a técnicos de dirección

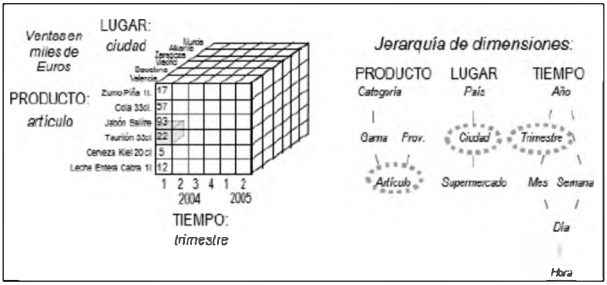
La *arquitectura* de un AD viene determinada por su situación central como fuente de información para las herramientas de análisis, tal y como se muestra en la primera ilustración de este capítulo (página 3).

Las *componentes* típicas de un almacén de datos pueden enumerarse como se indica a continuación:

- Sistema ETL (*Extraction, Transformation, Load*): realiza las funciones de *extracción* de las fuentes de datos (transaccionales o externas), *transformación* (limpieza, consolidación...) y la *carga* del AD, realizando:
    - extracción de los datos.
    - filtrado de los datos: limpieza, consolidación, etc.
    - carga inicial del almacén: ordenación, agregaciones, etc.
    - refresco del almacén: operación periódica que propaga los cambios de las fuentes externas al almacén de datos.
- Repositorio Propio de Datos: información relevante, metadatos.
- Interfaces y Gestores de Consulta: permiten acceder a los datos y sobre ellos se conectan herramientas más sofisticadas (OLAP, EIS, minería de datos).
- Sistemas de Integridad y Seguridad: se encargan de un mantenimiento global, copias de seguridad...

Las *herramientas de explotación* de los almacenes de datos han adoptado un **modelo multidimensional de datos**. En un esquema multidimensional se representa una actividad que es objeto de análisis (*hecho*) y las dimensiones que caracterizan la actividad (*dimensiones*). La información relevante sobre el hecho (*actividad*) se representa por un conjunto de indicadores (*medidas* o *atributos de hecho*). La información descriptiva de cada dimensión se representa por un conjunto de atributos (*atributos de dimensión*). Entre los atributos de una dimensión se definen *jerarquías*.

Se pueden obtener hechos a diferentes niveles de agregación. Es posible la obtención de medidas sobre los hechos parametrizadas por atributos de las dimensiones y restringidas por condiciones impuestas sobre las dimensiones. Un nivel de agregación para un conjunto de dimensiones se denomina *cubo*. El cubo que se muestra a continuación ilustra el hecho de las ventas en miles de euros de un artículo (PRODUCTO) en determinadas ciudades (LUGAR) en distintos momentos del tiempo (TIEMPO). De esta forma observamos la jerarquía de dimensiones PRODUCTO → LUGAR → TIEMPO.



La información de un almacén de datos se recopila en varios esquemas, cada uno de los cuales se denomina *datamart*. Los datamarts se definen para satisfacer las necesidades de un departamento o sección de la organización y contienen menos información de detalle y más información agregada. El almacén de datos puede estar formado por varios datamarts y, opcionalmente, por tablas adicionales.

El sistema encargado de la carga y mantenimiento del almacén de datos es el Sistema E.T.T. (Extracción - Transformación -Transporte). La construcción del Sistema E.T.T. es responsabilidad del equipo de desarrollo del almacén de datos. El Sistema E.T.T. es construido específicamente para cada almacén de datos. Aproximadamente 50% del esfuerzo. En la construcción del E.T.T. se pueden utilizar herramientas del mercado o programas diseñados específicamente.

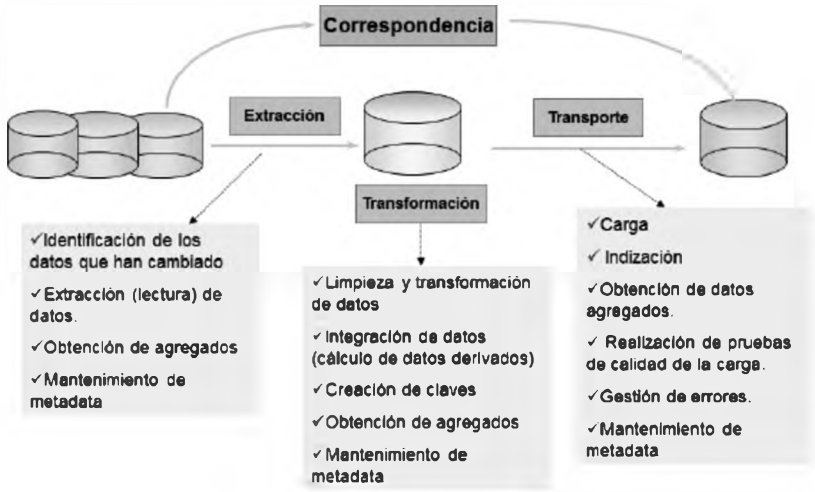
©Alfaomega-RC Libros

Las funciones del Sistema E.T.T. son la carga inicial (*initial load*) y el mantenimiento o *refresco* periódico: inmediato, diario, semanal, mensual... (*refreshment*).

El sistema E.T.T. es conocido también por E.T.L. (Extracción – Transformación – Load o carga).

En el proceso de carga y mantenimiento, a partir de fuentes externas de las que se realiza la extracción (que generalmente son bases de datos operacionales), se lleva a cabo el almacenamiento intermedio previo al transporte almacén de datos. El almacenamiento intermedio permite realizar transformaciones sin paralizar las bases de datos operacionales y el almacén de datos, almacenar metadatos y facilitar la integración de fuentes externas. El esquema siguiente ilustra estas fases:

El esquema que se presenta a continuación especifica las tareas a realizar en cada una de las fases del proceso de carga y mantenimiento de un almacén de datos.



HERRAMIENTAS OLAP, ROLAP Y MOLAP

Las herramientas de OLAP presentan al usuario una visión multidimensional de los datos (esquema multidimensional) para cada actividad que es objeto de análisis.

El usuario formula consultas a la herramienta OLAP seleccionando atributos de este esquema multidimensional sin conocer la estructura interna (esquema físico) del almacén de datos. Una consulta a un almacén de datos consiste generalmente en la obtención de medidas sobre los hechos parametrizadas por atributos de las dimensiones y restringidas por condiciones impuestas sobre las dimensiones. La herramienta OLAP genera la correspondiente consulta y la envía al gestor de consultas del sistema (p.ej. mediante una sentencia SELECT).

Por ejemplo, podemos plantearnos una consulta del tipo "Importe total de las ventas durante este año de los productos del departamento **bebidas**, por **trimestre** y por **categoría**". En este caso, las restricciones son: productos del departamento bebidas y ventas durante este año, mientras que los parámetros de la consulta **son:** por categoría de producto y por trimestre.

Se pueden presentar en forma tabular (relacional) los datos seleccionados asumiendo dos categorías en el departamento de *bebidas*: refrescos y zumos.

Categoría	Trimestre	Ventas
Refrescos	T1	2000000
Refrescos	T2	1000000
Refrescos	T3	3000000
Refrescos	T4	2000000
Zumos	T1	1000000
Zumos	T2	1500000
Zumos	T3	8000000
Zumos	T4	2400000

©Alfaomega-RC Libros

Pero también se puede realizar una presentación matricial (multidimensional) de los datos seleccionados.

trimestre / categoría	T1	T2	T3	T4
Refrescos	2000000	1000000	3000000	2000000
Zumos	1000000	1500000	8000000	2400000

Los parámetros de la consulta ("por trimestre" y "por categoría") determinan los criterios de agrupación de los datos seleccionados (ventas de productos del departamento *Bebidas* durante este año). La agrupación se realiza sobre dos dimensiones (Producto, Tiempo).

Pero lo interesante no es poder realizar consultas que, en cierto modo, se pueden hacer con selecciones, proyecciones, concatenaciones y agrupamientos tradicionales. Lo realmente interesante de las herramientas OLAP son sus operadores de refinamiento o manipulación de consultas como DRILL, ROLL, SLICE & DICE y PIVOT.

El carácter agregado de las consultas en el Análisis de Datos, aconseja la definición de nuevos operadores que faciliten la agregación (consolidación) y la disgregación (división) de los datos. Para la agregación tenemos el operador ROLL, que permite eliminar un criterio de agrupación en el análisis, agregando los grupos actuales. Para la desagregación tenemos el operador DRILL, que permite introducir un nuevo criterio de agrupación en el análisis, disgregando los grupos actuales.

Por ejemplo, en el esquema siguiente, mediante una operación DRILL, cada grupo (categoría-trimestre) de la consulta original se disgrega en dos nuevos grupos (categoría-trimestre-ciudad) para las ciudades de León y Valencia.

Categoría	Trimestre	Ventas
Refrescos	T1	2000000
Refrescos	T2	1000000
Refrescos	T3	3000000
Refrescos	T4	2000000
Zumos	T1	1000000
Zumos	T2	1500000
Zumos	T3	8000000
Zumos	T4	2400000

drill

Categoría	Trimestre	Ciudad	Ventas
Refrescos	T1	Valencia	1000000
Refrescos	T1	León	1000000
Refrescos	T2	Valencia	400000
Refrescos	T2	León	700000

A continuación se presenta un esquema de una operación de ROLL para la agregación de categorías de productos por trimestres y ventas.

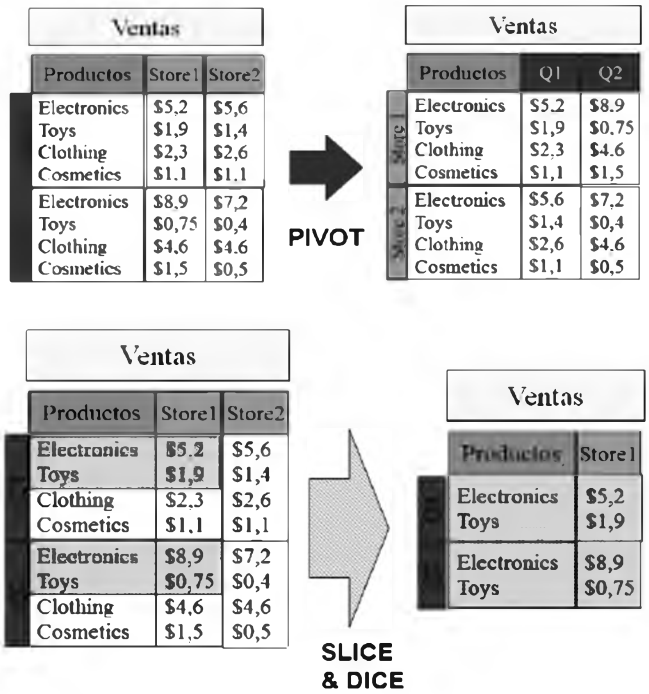
Categoría	Trimestre	Ventas
Refrescos	T1	2000000
Refrescos	T2	1000000
Refrescos	T3	3000000
Refrescos	T4	2000000
Zumos	T1	1000000
Zumos	T2	1500000
Zumos	T3	8000000
Zumos	T4	2400000

roll

Categoría	Ventas
Refrescos	8000000
Zumos	12900000

Existen otras operaciones de OLAP típicas como SLICE & DICE (para seleccionar y proyectar datos en el informe) y PIVOT (para reorientar las dimensiones en el informe).

©Alfaomega-RC Libros



Las herramientas de OLAP se caracterizan por ofrecer una visión multidimensional de los datos (matricial), no imponer restricciones sobre el número de dimensiones, ofrecer simetría para las dimensiones, permitir definir de forma flexible (sin limitaciones) sobre las dimensiones (restricciones, agregaciones y jerarquías entre ellas), ofrecer operadores intuitivos de manipulación (como drill-down, roll-up, slice-and-dice y pivot) y también por ser transparentes al tipo de tecnología que soporta el almacén de datos (ROLAP o MOLAP).

El almacén de datos y las herramientas OLAP se pueden basar físicamente en varias organizaciones:

- *Sistemas ROLAP*: se implementan sobre tecnología relacional, pero disponen de algunas facilidades para mejorar el rendimiento (índices de mapas de bits, índices de JOIN).

- *Sistemas MOLAP*: disponen de estructuras de almacenamiento específicas (arrays) y técnicas de compactación de datos que favorecen el rendimiento del almacén.

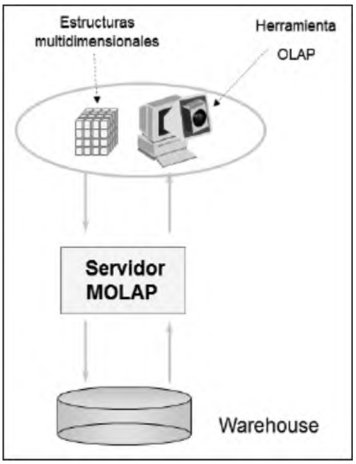
- *Sistemas HOLAP*: sistemas híbridos entre los dos anteriores.

En el caso de los Sistemas ROLAP, el almacén de datos se construye sobre un SGBD Relacional. Los fabricantes de SGBD relacionales ofrecen extensiones y herramientas para poder utilizar el SGBDR como un Sistema Gestor de Almacenes de Datos. Entre estas extensiones se encuentran habitualmente los índices de mapa de bits, los índices de JOIN, técnicas de particionamiento de los datos, optimizadores de consultas y extensiones del SQL (operador CUBE, ROLL-UP, etc.).

En el caso de los sistemas MOLAP, el propósito específico son las estructuras de datos (arrays) y las técnicas de compactación. El objetivo de los sistemas MOLAP es almacenar físicamente los datos en estructuras multidimensionales de forma que la representación externa y la representación interna coincidan.

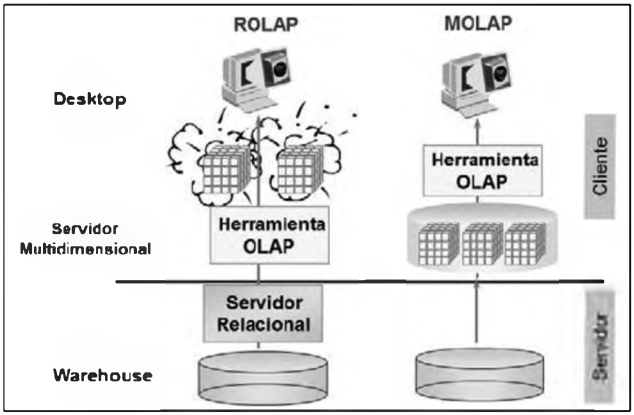
El servidor MOLAP construye y almacena datos en estructuras multidimensionales. La herramienta de OLAP presenta estas estructuras multidimensionales. Los datos suelen ser arrays extraídos del almacén de datos. El almacenamiento y los procesos son muy eficientes y la complejidad de la base de datos se oculta a los usuarios. Además, el análisis se hace sobre datos agregados y métricas o indicadores precalculados.

El esquema siguiente pretende relacionar toda esta tipología de conceptos.



El diagrama siguiente muestra una comparativa entre los sistemas ROLAP y MOLAP.

©Alfaomega-RC Libros



Los sistemas ROLAP pueden aprovechar la tecnología relacional, pueden utilizarse sistemas relacionales genéricos (más baratos o incluso gratuitos) y el diseño lógico corresponde al físico si se utiliza el diseño de Kimball.

Los sistemas MOLAP generalmente son más eficientes que los ROLAP y presentan un coste de los cambios en la visión de los datos y de la construcción de las estructuras multidimensionales.

INFORMES OLAP CON CUBE Y ROLLUP EN SQL

Cuando tenemos que mostrar información de nuestra base de datos a los usuarios habitualmente creamos informes con Visual Basic, con Crystal Reports, con Analysis Services de Microsoft SQL Server o con cualquier lenguaje relacional de bases de datos. Con estas herramientas podemos realizar cualquier tipo de cálculo sobre los datos almacenados, y especialmente calcular sumas, totales, promedios, etc.

Pero el lenguaje SQL también nos proporciona herramientas para hacer la mayor parte del trabajo en el servidor (ahorrándonos posteriores problemas). Existen cláusulas como GROUP BY para agrupar y las funciones de agregado para contar, sumar, promediar. ¿Y si queremos calcular subtotales y totales generales en una misma consulta? Pues también existen los operadores CUBE y ROLLUP que son los que vamos a tratar aquí. Ambos operadores suelen ser parte de la cláusula GROUP BY de la sentencia SELECT, tanto en ORACLE, como en TRANSACT SQL, etc.

Vamos a trabajar con la siguiente tabla por no complicar los ejemplos ni el código SQL. Además, una tabla como esta aunque no sea real sirve perfectamente para mostrar cómo funciona WITH CUBE.

Id	TipoTransaccion	Divisa	Cantidad
1	Entrada	Euro	200
2	Entrada	Euro	1300
3	Salida	Dolar	2000
4	Entrada	Libra	500
5	Salida	Dolar	1000
6	Entrada	Euro	300
7	Entrada	Dolar	5000
8	Salida	Libra	500
9	Entrada	Euro	700
10	Entrada	Libra	400
11	Entrada	Yen	20000
12	Salida	Libra	300
13	Entrada	Euro	4000
14	Salida	Yen	30000
15	Entrada	Libra	3000
16	Entrada	Euro	400
17	Entrada	Euro	900
18	Salida	Dolar	4000
19	Entrada	Dolar	1200
20	Salida	Libra	900
21	Entrada	Euro	2100
22	Entrada	Libra	200
23	Entrada	Yen	25000
24	Entrada	Libra	400
25	Entrada	Euro	700
26	Entrada	NULL	2000

Se trata de una tabla en la que se guarda información sobre transacciones económicas en las que tenemos tres tipos de datos. Si la transacción es de entrada o de salida, la moneda en la que se hace y la cantidad. Ahora vamos a empezar a hacer preguntas de tipo OLAP (*on-line analytical processing*) para analizar los datos que tenemos almacenados.

Podríamos preguntarnos: ¿cuántas transacciones tenemos de entrada y de salida? Para resolver esta cuestion podemos escribir una consulta sencilla con un GROUP BY y un COUNT:

```
SELECT TipoTransaccion, COUNT(IdTransaccion) Cantidad
        FROM Movimientos GROUP BY TipoTransaccion
TipoTransaccion                          Cantidad
---------------------------------------- -----------
Entrada                                  18
Salida                                   7
```


©Alfaomega-RC Libros

También podríamos preguntarnos: ¿qué divisa es la más usada? La respuesta es ahora un poco más elaborada, pero de nuevo nos basta con usar un GROUP BY. También utilizamos TOP 1 para quedarnos solo con el valor más alto después de ordenar la suma de manera descendente.

```
SELECT TOP 1 Divisa, SUM(Cantidad) Suma
          FROM Movimientos GROUP BY Divisa
          ORDER BY SUM(Cantidad) DESC
Divisa                                   Suma
---------------------------------------- -----------
Yen                                      75000
```

¿Y si queremos la cantidad de cada tipo de transacción en cada tipo de divisa? Ahora tenemos que agrupar por TipoTransaccion y Divisa para obtener la suma de las cantidades por esos conceptos. Podemos ver cuántos euros han salido o cuántos dólares han entrado.

```
SELECT   TipoTransaccion, Divisa, SUM(Cantidad) Cantidad
          FROM Movimientos GROUP BY TipoTransaccion, Divisa
          ORDER BY TipoTransaccion

TipoTransaccion              Divisa                 Cantidad
---------------------------- ------------------ -----------
Entrada                          Dolar                  6200
Entrada                          Euro                   10600
Entrada                          Libra                  4500
Entrada                          Yen                    45000
Salida                           Dolar                  7000
Salida                           Libra                  1700
Salida                           Yen                    30000
```

Pero podemos querer la información agrupada de más formas. Por ejemplo, para saber el total entrante, o el balance de yenes habrá que hacer cálculos adicionales o bien con nuevas consultas, o bien en nuestra aplicación cliente. Pero ahora tenemos WITH CUBE que nos permite crear nuevas dimensiones en nuestras consultas. Cuando usamos esta cláusula es como si estuviésemos haciendo a la vez todos los GROUP BY posibles y además mostrándolos en un único resultset. Añadamos WITH CUBE a la sentencia anterior.

```
SELECT   TipoTransaccion, Divisa, SUM(Cantidad) Cantidad
          FROM Movimientos GROUP BY TipoTransaccion, Divisa
          WITH CUBE

TipoTransaccion              Divisa             Cantidad
---------------------------- ------------------ -----------
Entrada                      Dolar              6200
Entrada                      Euro               10600
Entrada                      Libra              4500
Entrada                      Yen                45000
```

```
Entrada        NULL        66300
Salida         Dolar       7000
Salida         Libra       1700
Salida         Yen         30000
Salida         NULL        38700
NULL           NULL        105000
NULL           Dolar       13200
NULL           Euro        10600
NULL           Libra       6200
NULL           Yen         75000
```

Aquí se observan unos cuantos cambios. Primero vemos que sin necesidad de decirlo los datos se han ordenado por TipoTransaccion y dentro de TipoTransaccion por Divisa. Además, aparecen varios NULL por el medio de la tabla. Pero no os preocupéis que todo va bien y vamos a explicar este resultado con calma. Ahora cada fila es una de las posibles combinaciones de TipoTransaccion con Divisa, y las filas que contienen un NULL se tienen que leer pensando que donde está el NULL debería poner "todas". Es decir, la fila:

```
NULL           Dolar       13200
```

indica que la cantidad total de dólares tanto en entradas como en salidas (todos los TiposTransaccipn) es 13200. La fila:

```
Entrada        NULL        66300
```

indica que hay un valor total de 66300 para todas las entradas (es decir, para todas las divisas). Y por último la fila:

```
NULL           NULL        105000
```

nos indica que el total de movimientos (todos los TipoTransaccion) de entrada y salida en cualquier divisa (todas las divisas) es de 105000.

Como vemos el NULL representa un superagregado en la columna en la que está colocado. Este tipo de NULL no lo debemos confundir con un NULL normal. Ya sabemos que un NULL normal indica que desconocemos el valor mientras que este NULL indica una agrupación.

Vamos a insertar una nueva fila en nuestra tabla:

```
INSERT INTO Movimientos (TipoTransaccion, Divisa, Cantidad)
        VALUES ('Entrada', NULL, 2000)
```

¿Qué ocurre ahora si repetimos la consulta?

©Alfaomega-RC Libros

Hay una función llamada GROUPING que nos dice cuándo nuestro NULL es de verdad y cuándo no. Esta función nos devuelve un 1 si el nombre de la columna pasada como parámetro se usa como resumen y un 0 si no es así. Veamos un ejemplo:

```
SELECT   TipoTransaccion,
         Divisa, 'Todas las Divisas'=GROUPING(Divisa),
         SUM(Cantidad) Cantidad
         FROM Movimientos GROUP BY TipoTransaccion, Divisa
         WITH CUBE

TipoTransaccion          Divisa          Todas las Divisas    Cantidad
------------------------ --------------- -------------------- -----------
Entrada                  NULL            0                    2000
Entrada                  Dolar           0                    6200
Entrada                  Euro            0                    10600
Entrada                  Libra           0                    4500
Entrada                  Yen             0                    45000
Entrada                  NULL            1                    68300
Salida                   Dolar           0                    7000
Salida                   Libra           0                    1700
Salida                   Yen             0                    30000
Salida                   NULL            1                    38700
NULL                     NULL            1                    107000
NULL                     NULL            0                    2000
NULL                     Dolar           0                    13200
NULL                     Euro            0                    10600
NULL                     Libra           0                    6200
NULL                     Yen             0                    75000
```

Se observa que hay dos tipos de NULL en la columna de divisas. Las que corresponden al último registro que insertamos que tiene un NULL en divisa, y al que la función GROUPING le asocia un 0, y el NULL que podemos traducir por "Todas las divisas" al que la función GROUPING le asocia un 1.

Ahora mezclamos estas funciones nuevas con dos funciones conocidas, CASE e ISNULL, para darle un aspecto más elegante al resultado obtenido.

```
SELECT   TipoTransaccion,
         'Divisa' = CASE
                        WHEN GROUPING(Divisa)=1 THEN 'Todas'
                        ELSE ISNULL(Divisa, 'N/D')
                    END,
         SUM(Cantidad) Cantidad
         FROM Movimientos GROUP BY TipoTransaccion, Divisa
         WITH CUBE

TipoTransaccion      Divisa               Cantidad
-------------------- -------------------- -----------
Entrada              N/D                  2000
Entrada              Dolar                6200
Entrada              Euro                 10600
```

```
Entrada          Libra          4500
Entrada          Yen            45000
Entrada          Todas          68300
Salida           Dolar          7000
Salida           Libra          1700
Salida           Yen            30000
Salida           Todas          38700
NULL             Todas          107000
NULL             N/D            2000
NULL             Dolar          13200
NULL             Euro           10600
NULL             Libra          6200
NULL             Yen            75000
(donde pone "N/D" pues quiere decir no disponible)
```

Ahora vamos a escribir la consulta que nos devuelve toda la información que podemos pedir a los datos iniciales utilizando adecuadamente WITH CUBE.

```
SELECT    'TipoTransaccion' = CASE
                              WHEN GROUPING(TipoTransaccion)=1 THEN 'Todas'
                              ELSE ISNULL(TipoTransaccion, 'N/D')
                          END,
          'Divisa' = CASE
                              WHEN GROUPING(Divisa)=1 THEN 'Todas'
                              ELSE ISNULL(Divisa, 'N/D')
                          END,
          SUM(Cantidad) Cantidad
          FROM Movimientos GROUP BY TipoTransaccion, Divisa
          WITH CUBE

TipoTransaccion          Divisa          Cantidad
------------------------ --------------- -----------
Entrada                  N/D             2000
Entrada                  Dolar           6200
Entrada                  Euro            10600
Entrada                  Libra           4500
Entrada                  Yen             45000
Entrada                  Todas           68300
Salida                   Dolar           7000
Salida                   Libra           1700
Salida                   Yen             30000
Salida                   Todas           38700
Todas                    Todas           107000
Todas                    N/D             2000
Todas                    Dolar           13200
Todas                    Euro            10600
Todas                    Libra           6200
Todas                    Yen             75000
```

Mientras que WITH CUBE genera un conjunto de resultados que muestra agregados para todas las combinaciones de valores de las columnas seleccionadas, WHIT ROLLUP genera un conjunto de resultados que muestra agregados para una jerarquía de valores de las columnas seleccionadas.

©Alfaomega-RC Libros

Es decir, con CUBE aparecen los resultados totalizados por TipoTransaccion, por Divisa, y por totales absolutos, mientras que con ROLLUP solo aparecerían los totales agrupados por lo que nosotros indiquemos. Veámoslo agrupando por TipoTransaccion:

```
SELECT    'TipoTransaccion' = CASE
                              WHEN GROUPING(TipoTransaccion)=1 THEN 'Todas'
                              ELSE ISNULL(TipoTransaccion, 'N/D')
                    END,
          'Divisa' = CASE
                              WHEN GROUPING(Divisa)=1 THEN 'Todas'
                              ELSE ISNULL(Divisa, 'N/D')
                    END,
          SUM(Cantidad) Cantidad
          FROM Movimientos
          GROUP BY TipoTransaccion, Divisa
          WITH ROLLUP
```

TipoTransaccion	Divisa	Cantidad
Entrada	N/D	2000
Entrada	Dolar	6200
Entrada	Euro	10600
Entrada	Libra	4500
Entrada	Yen	45000
Entrada	Todas	68300
Salida	Dolar	7000
Salida	Libra	1700
Salida	Yen	30000
Salida	Todas	38700
Todas	Todas	107000

Obtenemos menos información que con WITH CUBE pero de manera más clara. Además, muchas veces con esto será suficiente.

Si la cantidad de datos a tratar es muy grande este tipo de consultas pueden consumir muchos recursos y tiempo, pero hay muchas soluciones para que esto no sea un problema. Por ejemplo cuando tenemos una consulta que va a resumir una serie de datos y la vamos a necesitar habitualmente podemos convertirla en una vista o guardar el resultado en una tabla, y así podemos recurrir al resultado sin perder tiempo volviendo a ejecutar la sentencia SQL.

Siguiendo con nuestro ejemplo podemos almacenar el resultado que obtuvimos con el CUBE en una tabla nueva con SELECT . . . INTO.

```
SELECT    'TipoTransaccion' = CASE
                              WHEN GROUPING(TipoTransaccion)=1 THEN 'Todas'
                              ELSE ISNULL(TipoTransaccion, 'N/D')
                    END,
          'Divisa' = CASE
```

```
                    WHEN GROUPING(Divisa)=1 THEN 'Todas'
                    ELSE ISNULL(Divisa, 'N/D')
            END,
        SUM(Cantidad) Cantidad
        INTO Resumen
        FROM Movimientos GROUP BY TipoTransaccion, Divisa
        WITH CUBE
select * from resultado
```

FUNCIONES DE CUBO EN EXCEL

Las **funciones de cubo en Excel** nos ayudan a obtener información de un cubo OLAP y colocar la información directamente en una hoja de Excel. De esta manera podemos combinar el potencial de las funciones de Excel y su motor de cálculo junto con los beneficios de un repositorio de datos multidimensional. Con las funciones de cubo podemos buscar datos de un cubo de OLAP como miembros, conjuntos, propiedades o valores y mezclarlos con otros cálculos y fórmulas de Excel. La tabla siguiente muestra las funciones de cubo de Excel.

FUNCIÓN	INGLÉS	DESCRIPCIÓN
CONJUNTOCUBO	CUBESET	Define un conjunto de miembros o tuplas calculado enviando una expresión establecida al cubo del servidor, que crea el conjunto y lo devuelve a Microsoft Excel. Obtiene el valor máximo de una columna entre los registros que cumplen con los criterios establecidos Sintaxis: *CONJUNTOCUBO(conexión, expresión_conjunto, [título], [criterio_ordenación], [ordenar_por])* • conexión (obligatorio): La cadena de conexión al cubo. • expresión_conjunto (obligatorio): La expresión MDX que nos ayudará a obtener el conjunto de miembros. • título (opcional): Texto que será mostrado como el título del cubo. • criterio_ordenación (opcional): Tipo de ordenación que se dará a los datos. 0 = Ninguno, 1 = Ascendente, 2 = Descendente. • ordenar_por (opcional): Valor por el cual deseamos ordenar los datos. Ejemplo: **CONJUNTOCUBO**("Finanzas","Order([Product].[Product].[Product Category].Members,[Measures].[Unit Sales],ASC)")

©Alfaomega-RC Libros

MIEMBROCUBO	CUBEMEMBER	Devuelve un miembro de un cubo OLAP. Sintaxis: *MIEMBROCUBO(conexión, expresión_conjunto, [título])* • conexión (obligatorio): La cadena de conexión al cubo. • expresión_conjunto (obligatorio): La expresión MDX que nos ayudará a obtener el miembro del cubo. • título (opcional): Texto que será mostrado como el título del cubo. Ejemplo: **MIEMBROCUBO**("Ventas","[Time].[Fiscal].[2010]")
MIEMBROKPICUBO	CUBEKPIMEMBER	Devuelve una propiedad de indicador clave de rendimiento (KPI) y muestra el nombre de KPI en la celda. Sintaxis: *MIEMBROKPICUBO(conexión, nombre_kpi, propiedad_kpi, [título])* • conexión (obligatorio): La cadena de conexión al cubo. • nombre_kpi (obligatorio): Nombre del KPI en el cubo. • propiedad_kpi (obligatorio): Componente KPI devuelto. 1 = Valor real, 2 = Valor de destino. • título (opcional): Texto que será mostrado en lugar de nompre_kpi y propiedad kpi. Ejemplo: **MIEMBROKPICUBO**("Ventas","MiKPIVentas",1)
MIEMBRORANGOCUBO	CUBERANKEDMEMBER	Devuelve el miembro Nth u ordenado de un conjunto. Sintaxis: *MIEMBRORANGOCUBO(conexión, expresión_conjunto, clasificación, [título])* • conexión (obligatorio): La cadena de conexión al cubo. • expresión_conjunto (obligatorio): Cadena de texto con la expresión de conjunto. • clasificación (obligatorio): Valor superior que se debe devolver. 1 = Valor superior, 2 = Segundo más alto, etc. • título (opcional): Texto que será mostrado como el título del cubo.

		Ejemplo: **MIEMBRORANGOCUBO**("Ventas",A1,1)
PROPIEDADMIEMBROCUBO	CUBEMEMBERPROPERTY	Devuelve el valor de una propiedad de miembro en el cubo. Sintaxis: *PROPIEDADMIEMBROCUBO(conexión, expresión_miembro, propiedad)* • conexión (obligatorio): La cadena de conexión al cubo. • expresión_miembro (obligatorio): Expresión multidimensional (MDX) de un miembro dentro del cubo. • propiedad (obligatorio): Nombre de la propiedad a ser devuelta. Ejemplo: **PROPIEDADMIEMBROCUBO**("Ventas","[Time].[Fiscal].[2010]",A1)
RECUENTOCONJUNTOCUBO	CUBESETCOUNT	Devuelve el número de elementos de un conjunto. Sintaxis: *RECUENTOCONJUNTOCUBO(conjunto)* • conjunto (obligatorio): Cadena de texto que se evalúa como un conjunto. Se puede utilizar la función CONJUNTOCUBO. Ejemplo: **RECUENTOCONJUNTOCUBO**(CONJUNTOCUBO("Ventas","[Product].[All Products].Children","Products",1,"[Measures].[Sales Amount]"))
VALORCUBO	CUBEVALUE	Devuelve un valor agregado del cubo. Sintaxis: *VALORCUBO(conexión, [expresión_miembro1], [expresión_miembro2], ...)* • conexión (obligatorio): La cadena de conexión al cubo. • expresión_miembro (opcional): Expresión multidimensional (MDX) que se evalúa como un miembro o tupla dentro del cubo. Ejemplo: **VALORCUBO**("Ventas","[Measures].[Profit]","[Time].[2010]")

©Alfaomega-RC Libros

CAPÍTULO 2

MICROSOFT Y LA INTELIGENCIA DE NEGOCIOS. POWER BI. INFORMES CON POWER QUERY Y POWER PIVOT

INTRODUCCIÓN Y COMPONENTES DE POWER BI

El crecimiento de datos aumenta cada día, y con él, la necesidad de inteligencia empresarial con características de autoservicio, capaz de convertir los datos en perspectivas con todo lo que necesita para tomar decisiones. Pero a menudo, las herramientas necesarias para extraer las perspectivas son tan difíciles de manejar como los propios datos. Microsoft Power BI para Office 365 es una colección de nuevas funciones y servicios que, juntos, permiten visualizar datos, compartir hallazgos y trabajar con nuevas formas muy intuitivas. Power BI ayuda a ampliar los conocimientos de Excel, así como su capacidad para entender y aprovechar el mundo de los datos, valiosos y en constante crecimiento, pero a veces inmanejables.

Excel 2013, con su colección de nuevas características y complementos, proporciona un entorno familiar para crear y compartir un atractivo contenido de BI. Se puede usar Excel para crear contenido e informes atractivos, y ampliar los conocimientos con Power BI para Office 365, con uso compartido interactivo, colaboración y administración de datos. Cada nueva característica es atractiva por sí sola, pero el valor de Power BI se encuentra en cómo colaboran entre sí sus distintas características, y cómo las perspectivas que se obtienen se comparten de forma interactiva, pero con seguridad.

A cualquier usuario de Excel que necesite las funcionalidades de BI con características de autoservicio le resultará muy útil Power BI. Con su perfecta interacción con Excel, se puede acceder a Power BI inmediatamente, y se conecta de forma intuitiva al software y los servicios que ya usan los creadores de informes y los usuarios avanzados de Excel.

Creadores de informes/usuarios, administradores de datos, profesionales de TI y consumidores de informes pueden beneficiarse de las ventajas que ofrecen las características de Power BI. Un creador de informes como, por ejemplo, un analista de datos, un consultor de datos o BI o un usuario estándar de Excel, probablemente usará las características de uso compartido, colaboración y búsqueda de Power BI a menudo y con todas sus funciones. Los servicios en línea que admiten colaboración, como la distribución de informes interactivos y el uso compartido de libros, son un área de trabajo central y un hub de información de gran importancia. A un administrador de datos como, por ejemplo, un científico que trabaja con datos, un administrador de datos o un profesional de TI, las características de servicios en línea de Power BI le permiten ofrecer un acceso seguro y específico a los recursos de datos. También puede certificar fuentes o consultas de datos y, con ello, identificar (y diferenciar) esos elementos concediéndoles su sello de aprobación. También son importantes los elementos relacionados con los servicios en línea de algunas herramientas de BI con características de autoservicio como, por ejemplo, la identificación, selección y distribución seguras de fuentes de datos locales y públicas.

Excel permite crear contenido (como libros, modelos de datos y visualizaciones) que se puede publicar y compartir en Power BI para Office 365. Las herramientas de cliente de Power BI están ahora integradas en Excel. En concreto son:

- **Power Query**: Permite detectar los datos de los orígenes de datos públicos y corporativos y conectarse a ellos fácilmente. Incluye nuevas funcionalidades de búsqueda de datos, así como funcionalidades para transformar y combinar fácilmente los datos de varios orígenes de datos para poder seguir analizándolos en Excel.

- **Power Pivot**: Permite crear sofisticados modelos de datos en Excel mediante relaciones, medidas personalizadas, jerarquías y KPI. Los modelos de Power Pivot se ejecutan en la memoria para que los usuarios puedan analizar cientos de millones de filas de datos con un rendimiento espectacular.

©Alfaomega-RC Libros

- **Power View**: Permite crear fácilmente informes y vistas analíticas con gráficos interactivos que ayudan a explorar y presentar los datos de forma visual en Excel.

- **Power Map**: Permite explorar y navegar por datos geoespaciales mediante mapas 3D dentro de Excel.

Con las funciones de BI implementadas sobre Excel 2013, es fácil detectar y visualizar los datos. Estas cuatro características (Power Query, Power Pivot, Power View y Power Map) colaboran entre sí a la perfección. Con Power Query, puede buscar datos, tanto públicos como privados, en toda la nube. Si usa varios orígenes de datos, puede filtrar los datos, darles forma, combinarlos y anexarlos con Power Query sin tener que incluirlos en Excel. Puede ver los datos, darles forma y enviarlos directamente al modelo de datos de Power Pivot. Posteriormente, puede crear cálculos y campos nuevos en Power Pivot para obtener solo los conjuntos de datos que quiera. Power View usa el modelo de datos para dar vida a esos datos. Es posible crear informes y dejar que otros usuarios interactúen con ellos y los exploren en profundidad. Las visualizaciones cambiarán a medida que seleccionan los elementos del informe.

USUARIOS DE POWER BI

A cualquier usuario de Excel que necesite las funcionalidades de BI con características de autoservicio le resultará muy útil Power BI. Con su perfecta interacción con Excel, se puede acceder a Power BI inmediatamente, y se conecta de forma intuitiva al software y los servicios que ya usan los creadores de informes y los usuarios avanzados de Excel. Tanto el usuario o creador de informes, como otras personas de distintos roles, también pueden beneficiarse de las ventajas que ofrecen las características de Power BI. Estos roles se dividen en tres categorías: creadores de informes/usuarios, administradores de datos y profesionales de TI, y consumidores de informes. Vamos a verlos uno a uno.

Un creador de informes como, por ejemplo, un analista de datos, un consultor de datos o BI o un usuario estándar de Excel, probablemente usará las características de uso compartido, colaboración y búsqueda de Power BI a menudo y con todas sus funciones. Los servicios en línea que admiten colaboración, como la distribución de informes interactivos y el uso compartido de libros, son un área de trabajo central y un hub de información de gran importancia.

Un administrador de datos como, por ejemplo, un científico que trabaja con datos, un administrador de datos o un profesional de TI, las características de servicios en línea de Power BI le permiten ofrecer un acceso seguro y específico a los recursos de datos. También puede certificar fuentes o consultas de datos y, con ello, identificar (y diferenciar) esos elementos concediéndoles su sello de aprobación. También son importantes los elementos relacionados con los servicios en línea de algunas herramientas de BI con características de autoservicio como, por ejemplo, la identificación, selección y distribución seguras de fuentes de datos locales y públicas. Si está en esta categoría, consulte la Guía de aprovisionamiento de Power BI, que muestra cómo hacer que Power BI se ejecute en la organización.

Un consumidor de informes, desde un estudiante hasta un director general, puede colaborar en informes, compartirlos con seguridad e interactuar con ellos mediante Power BI para tomar decisiones más inteligentes, ágiles y fundadas. Esto incluye el uso de un hub en línea, integrado en Office 365, diseñado específicamente para la colaboración en BI.

POWER QUERY

Power Query permite a los usuarios buscar datos de todo tipo de fuentes y tipos, combinarlos, refinarlos, darles forma y descargarlos. En la pestaña Power Query pueden realizarse fácilmente búsquedas. Para ello se hace clic sobre el complemento POWERQUERY de Excel y se elige la pestaña *Búsqueda online* (Figura 2-1).

Aparece el panel *Búsqueda en línea*. Empezaremos con una búsqueda del S&P 500, que se considera una muestra bastante representativa de la bolsa. El objetivo es que, si consigue encontrar una lista y los datos asociados, se podrán crear informes interesantes y representativos.

Sorprendentemente, en el panel *Búsqueda en línea* aparecen todo tipo de resultados. El primer resultado de su búsqueda en línea es un listado de los componentes del S&P 500 de una página web de Wikipedia (Figura 2-2).

©Alfaomega-RC Libros

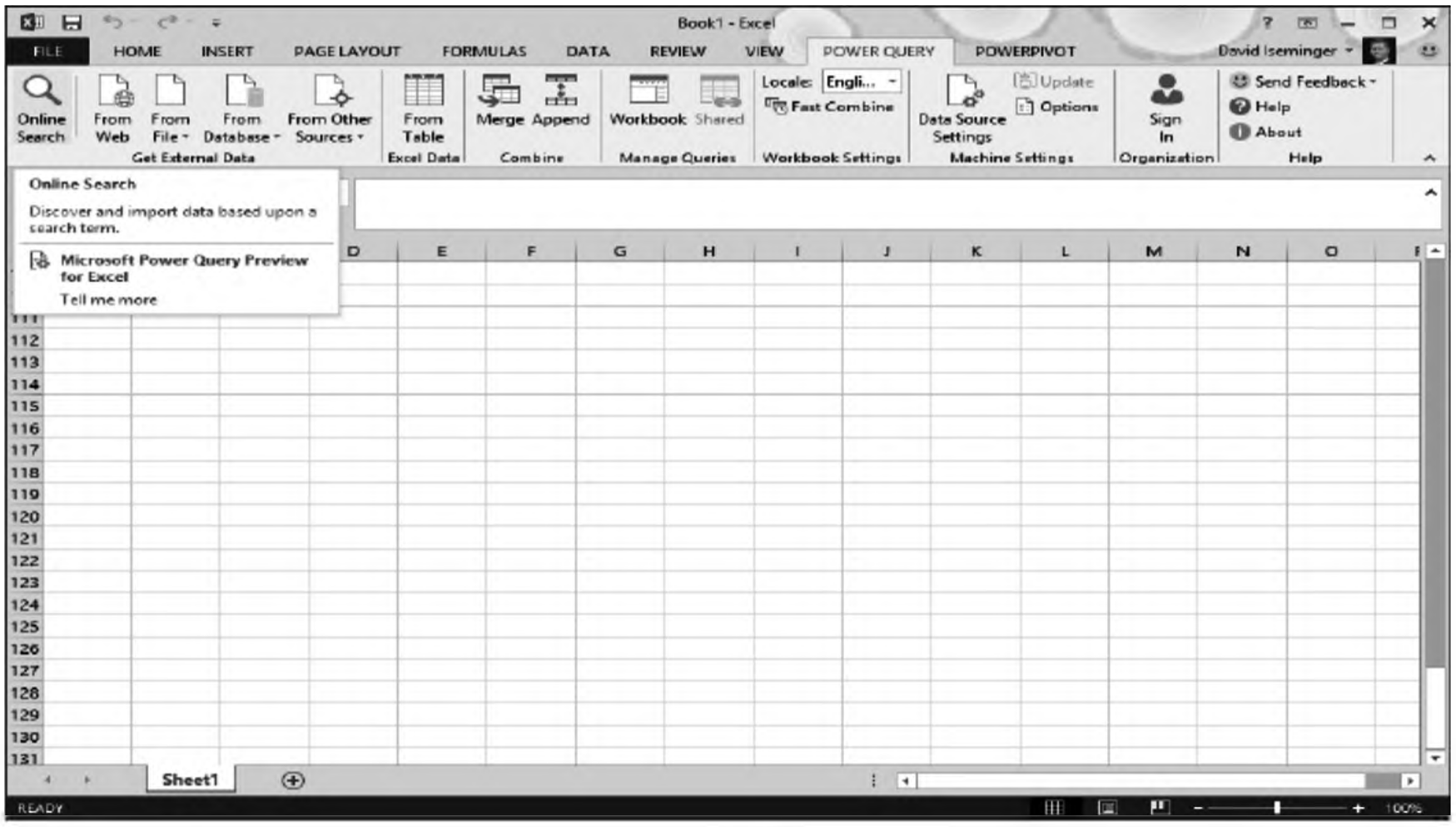


Figura 2-1

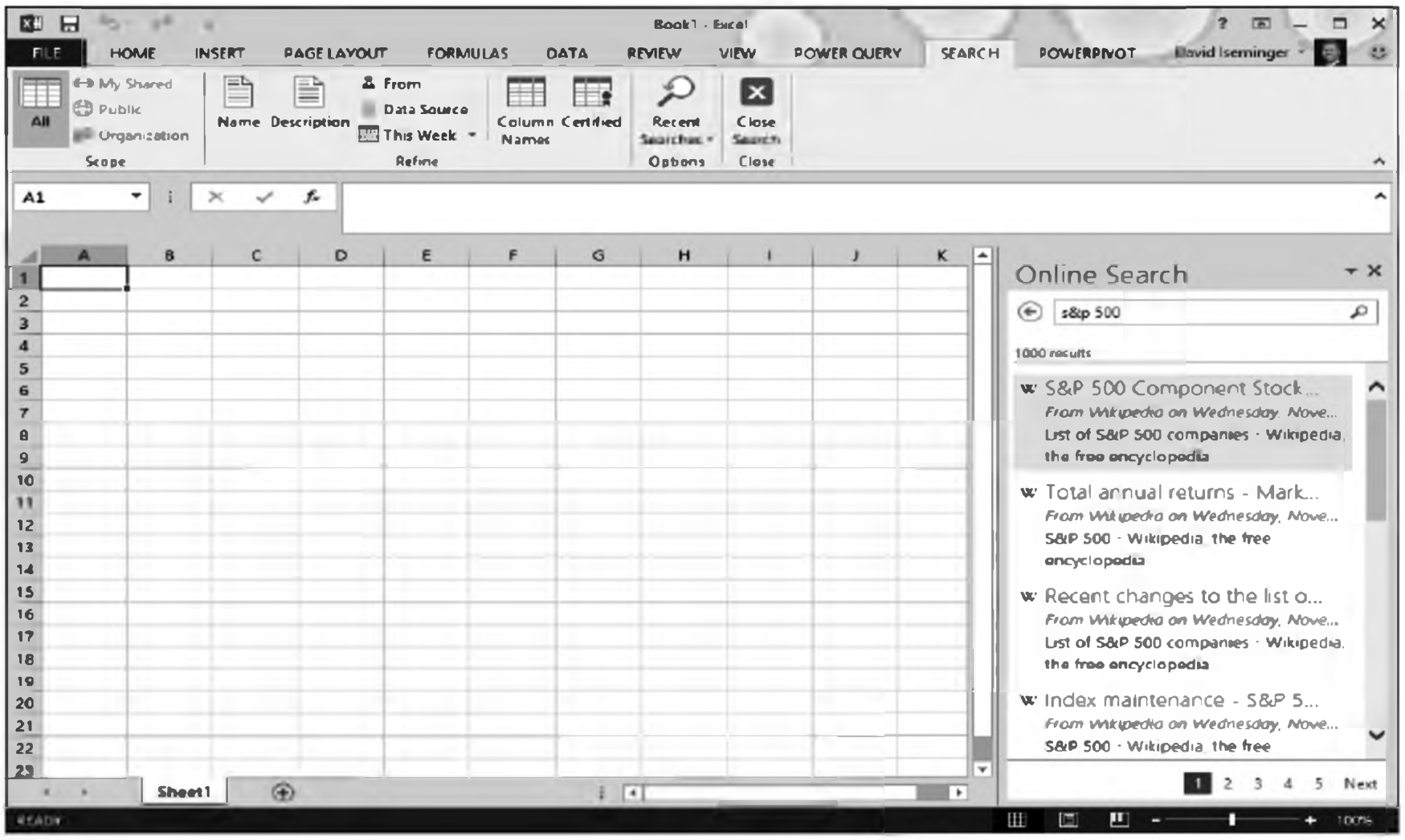


Figura 2-2

Al mantener el ratón sobre ese primer resultado de la búsqueda, se presenta información sobre ese origen de datos. Los términos de búsqueda que escribió aparecen resaltados (Figura 2-3).

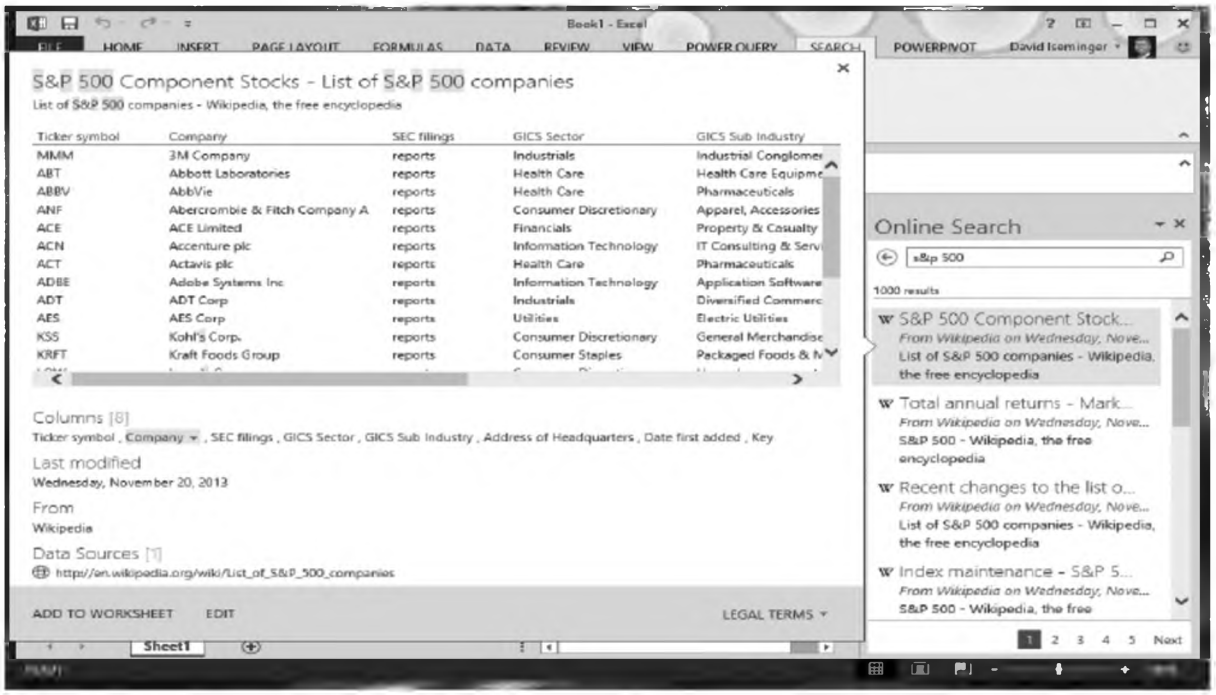


Figura 2-3

Ahora se selecciona *Agregar a hoja de cálculo* en la parte inferior del panel de información emergente, lo que le permite descargar los datos a Excel.

Ahora se necesitan datos para mezclar con la lista del S&P 500. Se buscarán datos de las cotizaciones diarias de hace unos años, durante la crisis, y se cargarán en Power Query para darles forma y verlos. Esta tarea se realizará desde la pestaña *Archivo* de Excel, eligiendo el origen de datos adecuado (Figura 2-4).

Power Query presentará la ventana *Editor de consultas*, que muestra todos los orígenes de datos disponibles en esa carpeta (Figura 2-5). Podría cargar todos los datos de esa carpeta haciendo clic en el icono de flecha doble hacia abajo que hay a la derecha de la columna Contenido. Pero ahora quiere obtener los libros de Excel uno por uno.

El nombre predeterminado de la consulta es Consulta1, como se ve en el panel *Configuración de consulta* que hay a la derecha, para ello cambia el nombre para que sea representativo de los datos: NYSE 2009 - diario. Cambiar el nombre le será útil cuando quiera usar la consulta con otros orígenes de datos, o cuando quiera compartir la consulta con otras personas.

©Alfaomega-RC Libros

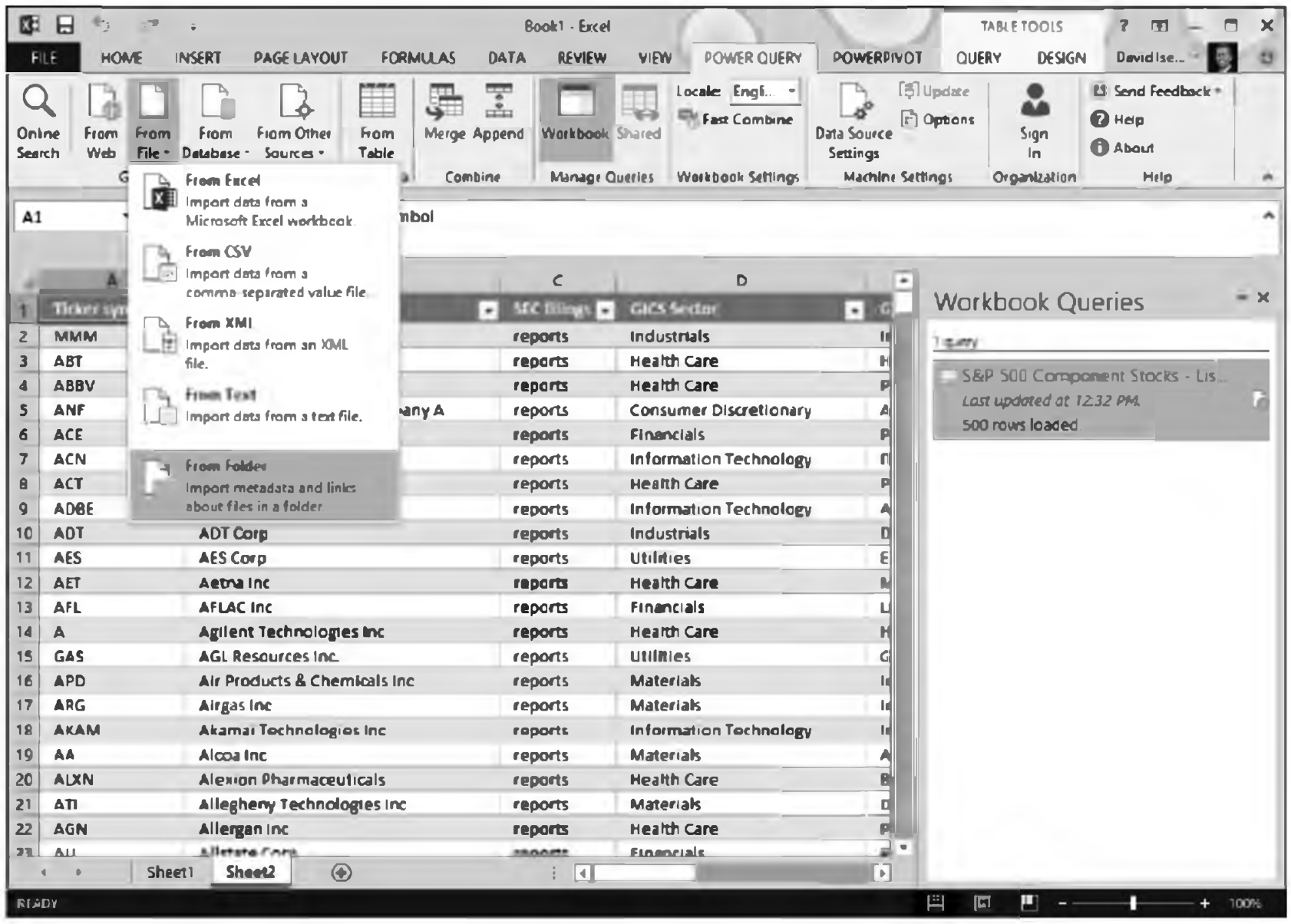


Figura 2-4

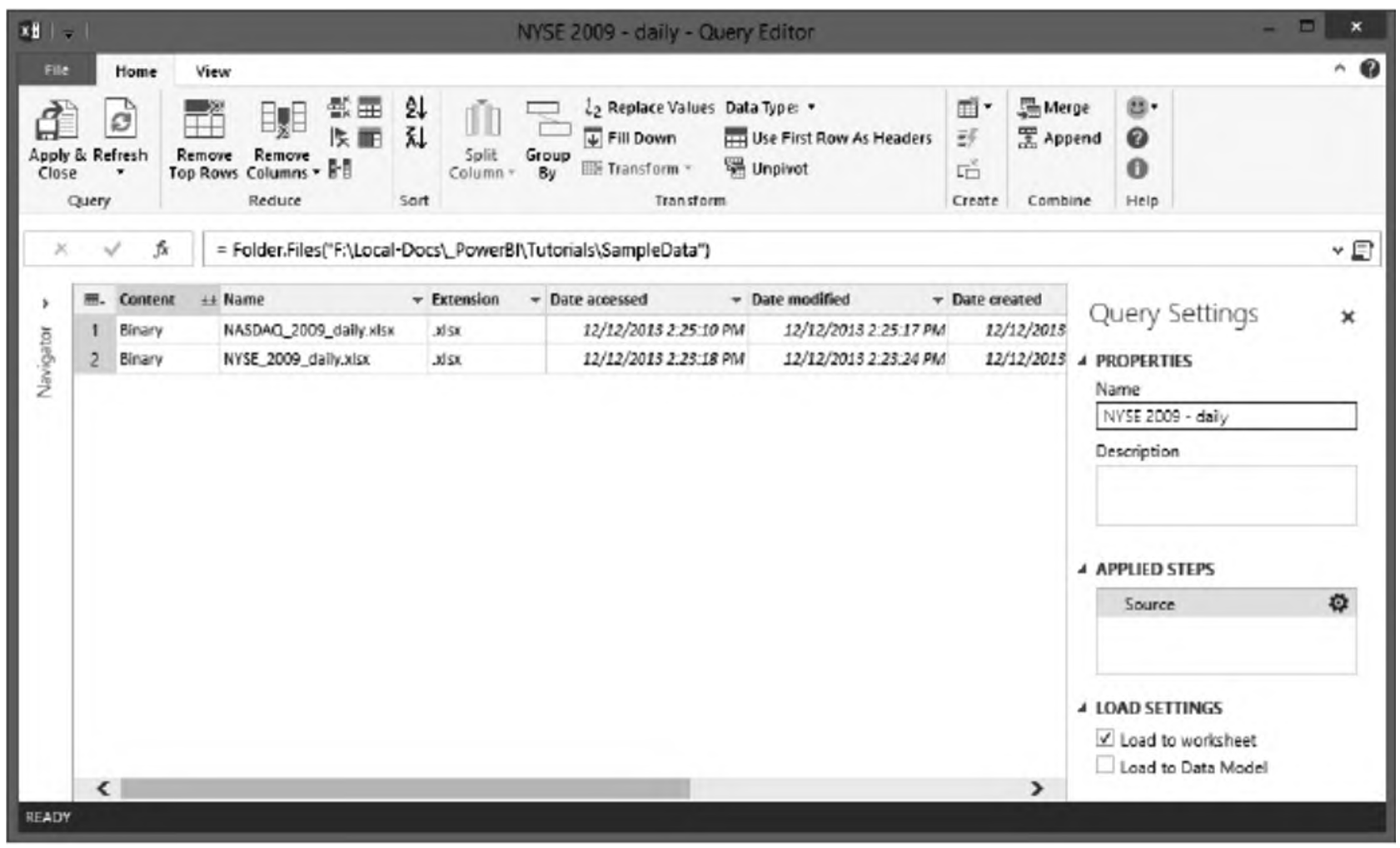


Figura 2-5

Habitualmente, los conjuntos de datos ocupan mucho, y no será conveniente descargar los datos al libro. Bastará con ver las columnas y decidir cómo filtrar los datos y darles forma. Luego se bajará solamente el subconjunto que va a usar. En *Configuración de carga*, desactive las casillas *Cargar en hoja de cálculo* y *Cargar en modelo de datos* (Figura 2-6).

Figura 2-6

Se observa que Power Query describe el origen de datos, en este caso un archivo de Excel, como Binario. Si quiere profundizar en los detalles de los datos, haga clic con el botón secundario en *Binario* (Figura 2-7) y elija *Explorar en profundidad* (al hacer clic o pulsar en la palabra Binario también se exploran los datos en profundidad). Excel muestra los detalles del libro de Excel. En este caso, contiene una tabla (Figura 2-8).

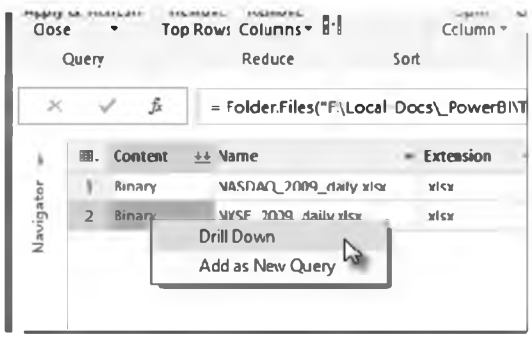


Figura 2-7

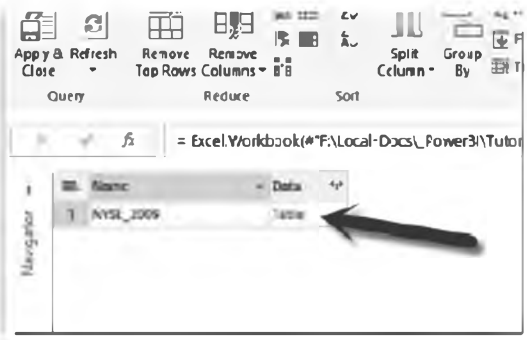


Figura 2-8

©Alfaomega-RC Libros

Haciendo clic en *Tabla* se obtienen los datos con los encabezados de columna en la primera fila (Figura 2-9), en lugar de mostrarse como encabezados de campo. Haciendo clic en *Usar primera fila como encabezados* en la cinta de opciones, Power Query usa la primera fila como encabezados (Figura 2-10). Después haga clic en *Aplicar y cerrar* en la sección *Consulta* de la cinta de opciones.

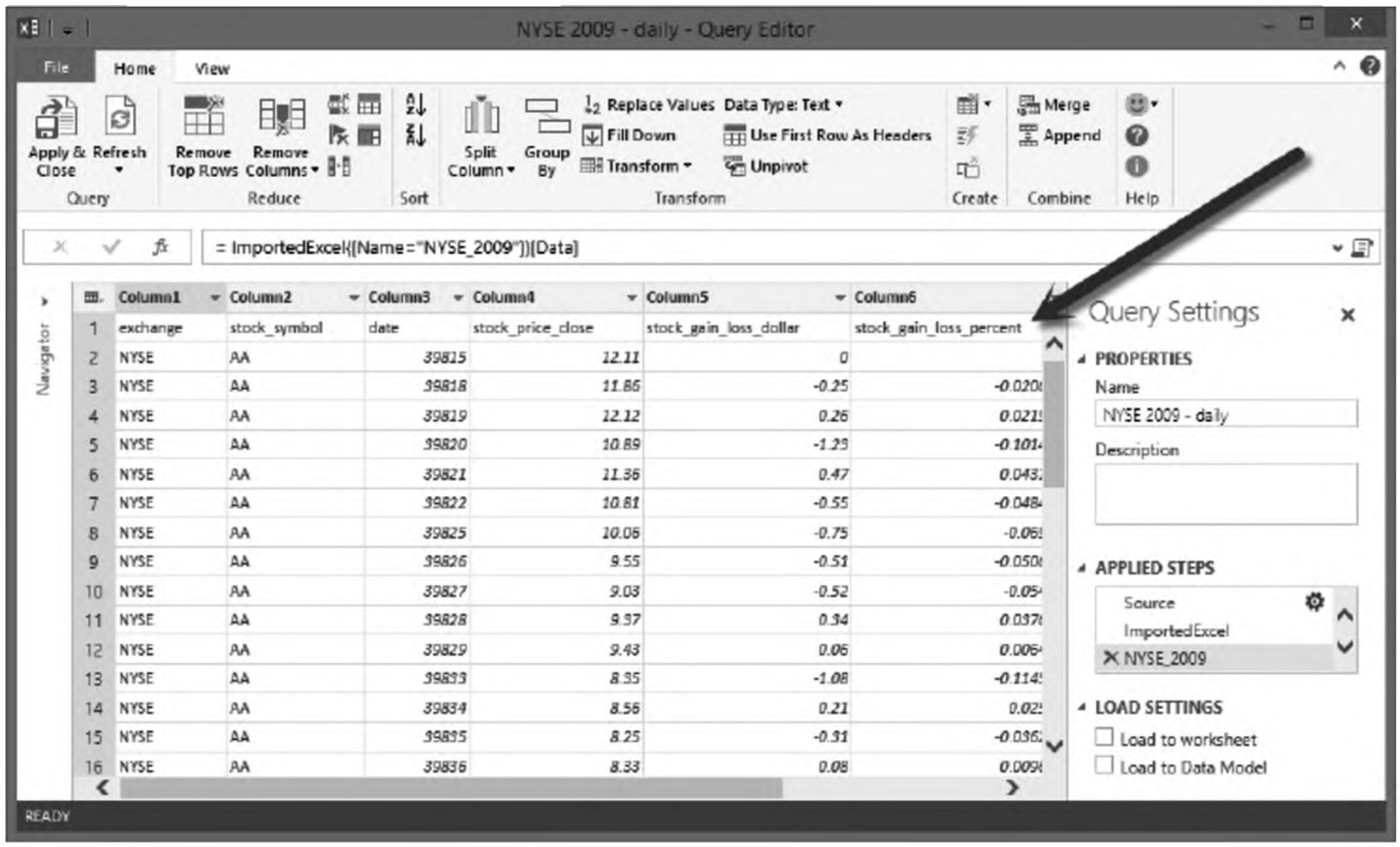


Figura 2-9

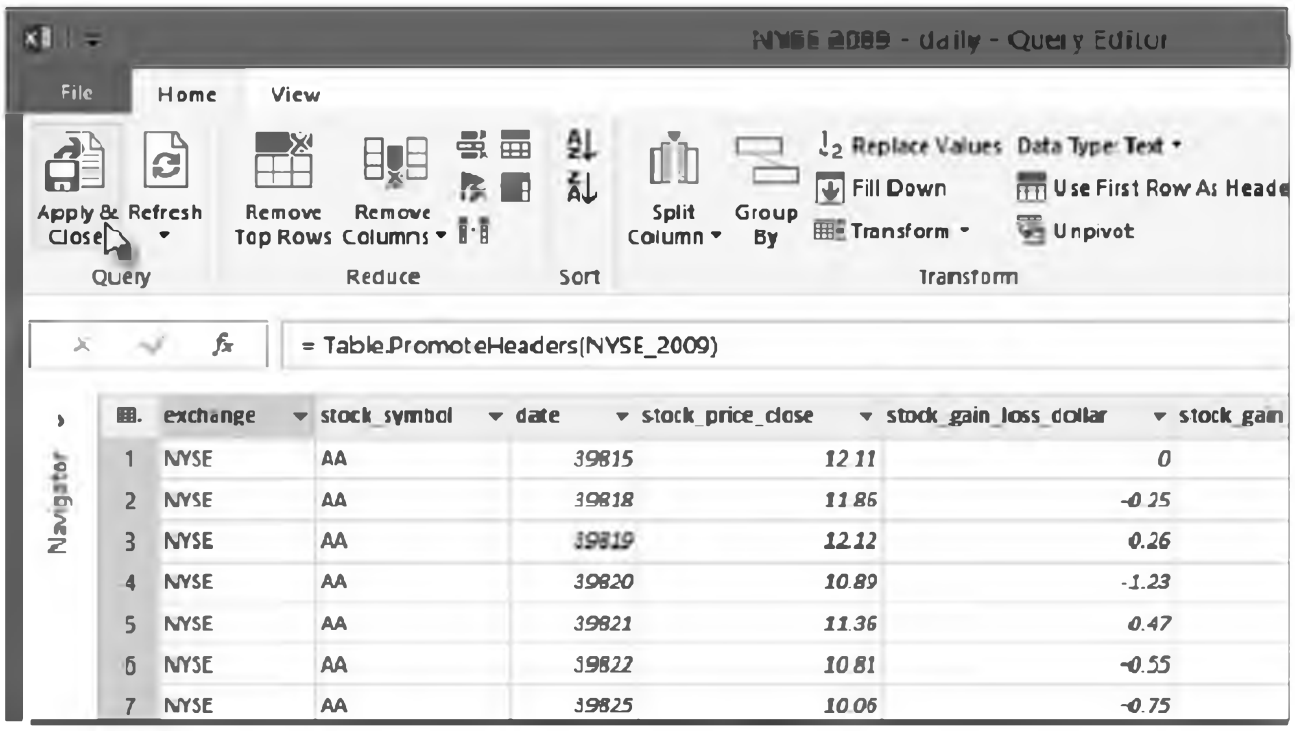


Figura 2-10

En Excel, la consulta está disponible en el panel *Consultas del libro*. Como desactivó todas las casillas de la sección *Configuración de carga* en la ventana *Editor de consultas* al crear la consulta, la consulta NYSE 2009 - diario se muestra en el panel *Consultas del libro*, con la carga deshabilitada (Figura 2-11).

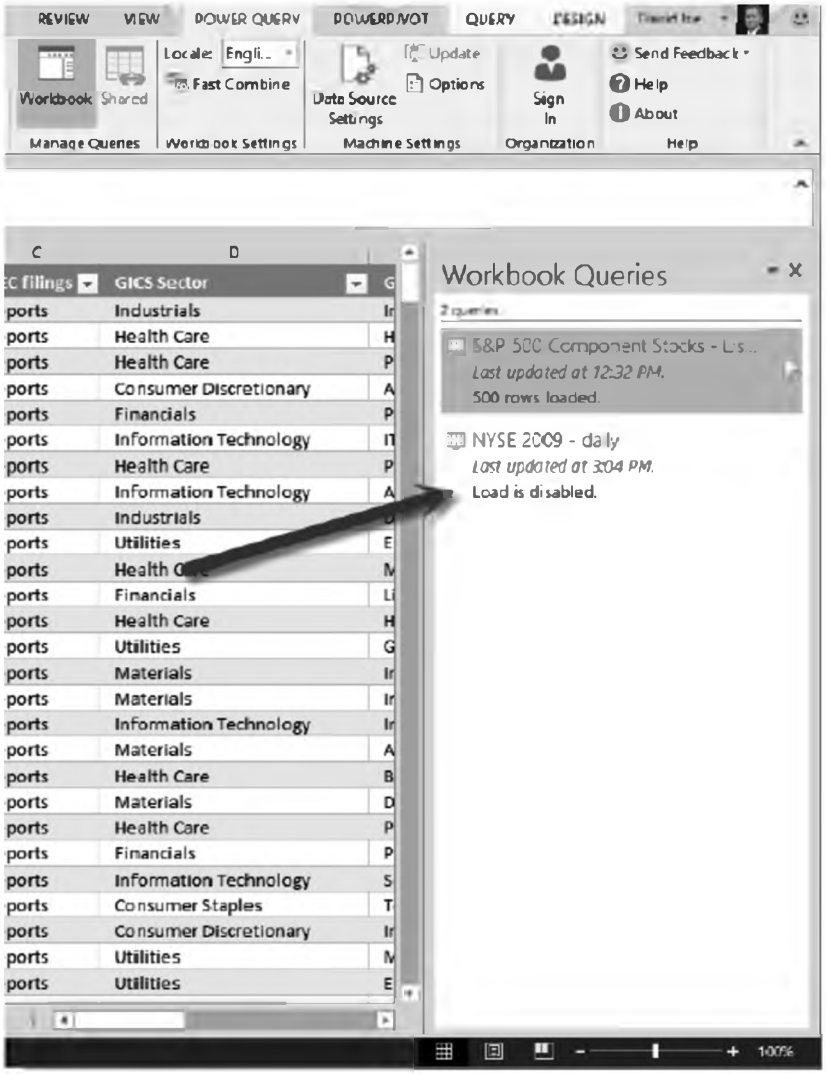


Figura 2-11

Ahora que tenemos los datos de las cotizaciones diarias y los datos del S&P 500, podemos combinarlos. Al hacerlo, se incluyen solamente los datos que necesita. En este caso, solo necesita los datos diarios de los símbolos de valores correspondientes a los símbolos bursátiles que forman parte del S&P 500.

Usamos Power Query para combinar los datos a través de Merge (Figura 2-12).

©Alfaomega-RC Libros

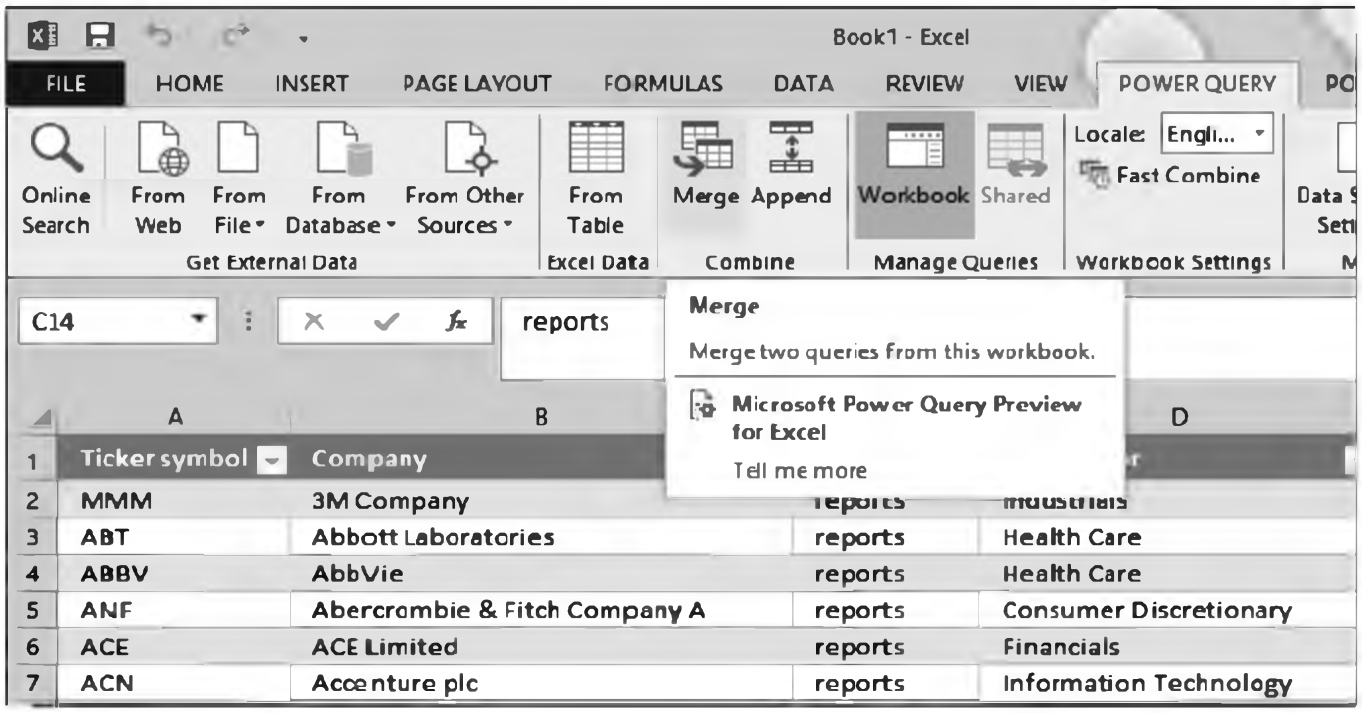


Figura 2-12

Power Query le permite elegir la tabla principal y las columnas que coinciden entre sí, de entre las consultas disponibles en el libro. En este caso, las columnas coincidentes son los símbolos de los valores, por lo que debe seleccionar la columna Símbolo del valor de la tabla del S&P 500 y símbolo_valor de la tabla del NYSE (Figura 2-13).

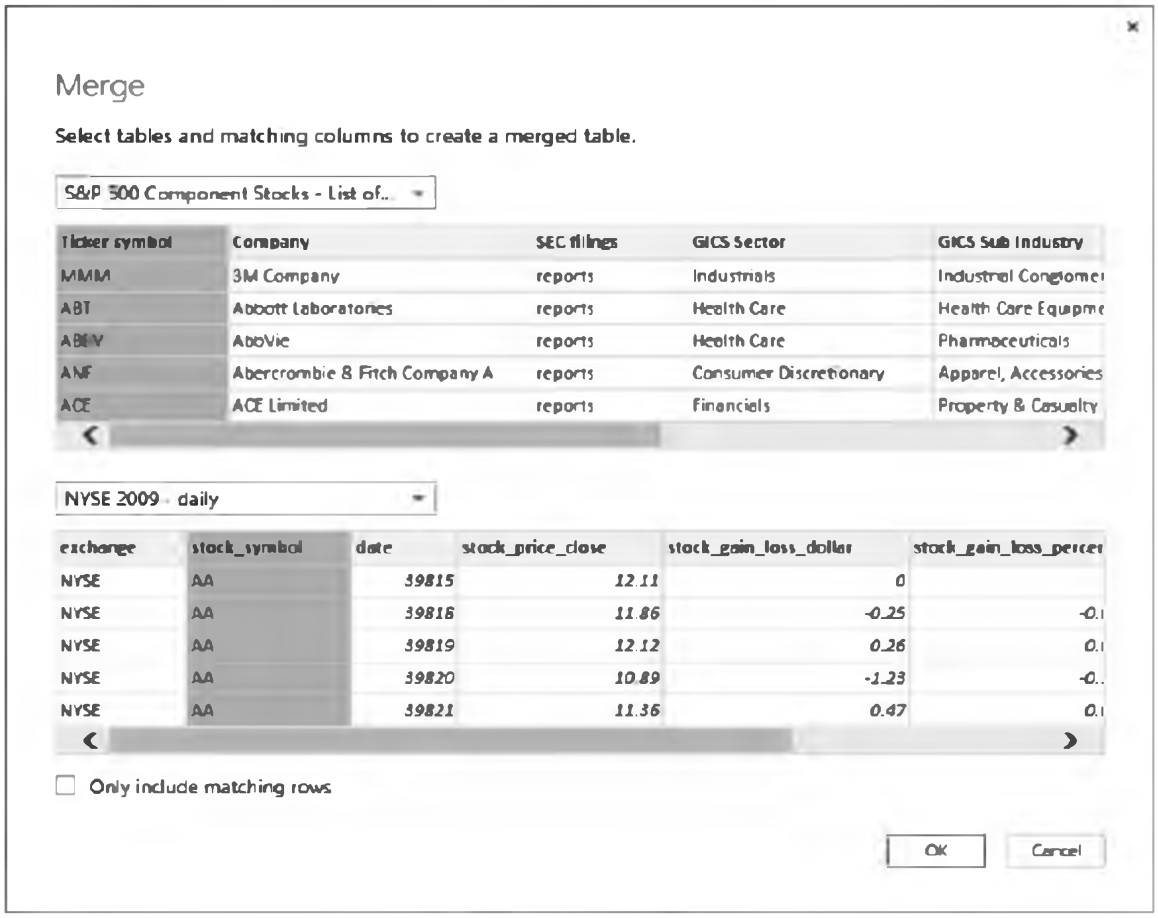


Figura 2-13

Al hacer clic en *Aceptar*, se le solicita que especifique la configuración de privacidad de los datos de cada conjunto de datos (Figura 2-14). Selecciona las opciones correspondientes y haga clic en *Guardar*.

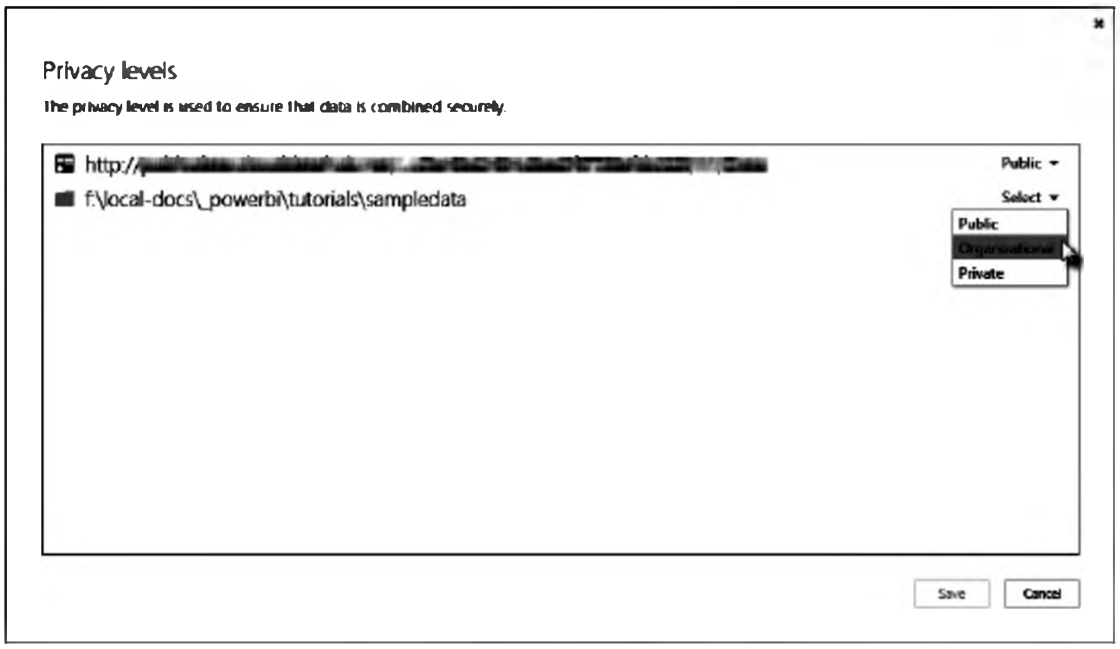


Figura 2-14

Después, se cambia el nombre de la consulta de Combinación1 a Combinación SP500-NYSE. En la ventana Editor de consultas, se puede dar forma a los datos según lo que se necesita. Por ejemplo, no necesita la columna que describe los archivos de informes de SEC, así que se quita (Figura 2-15).

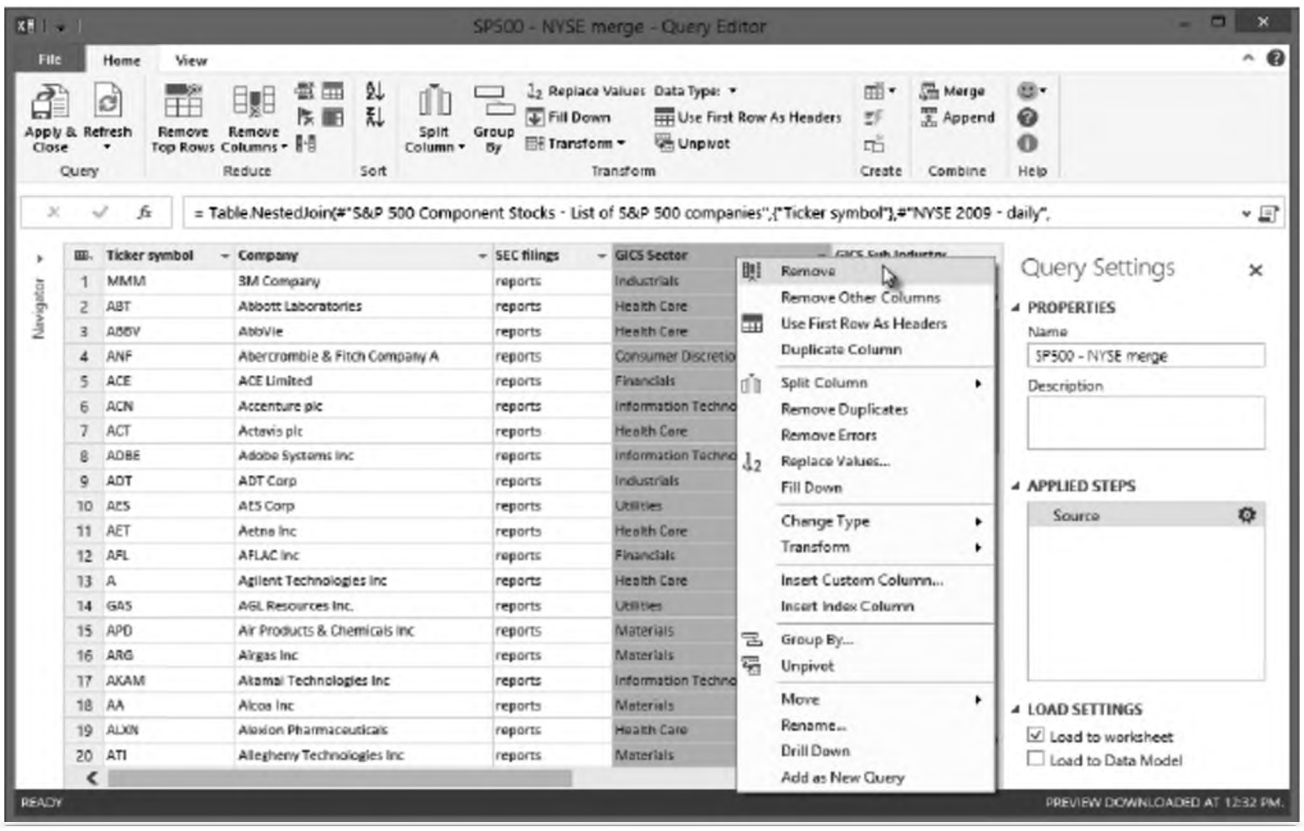


Figura 2-15

©Alfaomega-RC Libros

La tabla con la que se combinaron los datos del S&P 500 aparece como columnas expandibles, que se muestran en la parte final de las columnas disponibles. Cuando se hace clic en el icono de flecha doble del lado derecho del encabezado NuevaColumna, aparecen todas las columnas de la combinación NYSE 2009 - diario, disponibles para que ella las incluya si quiere (Figura 2-16).

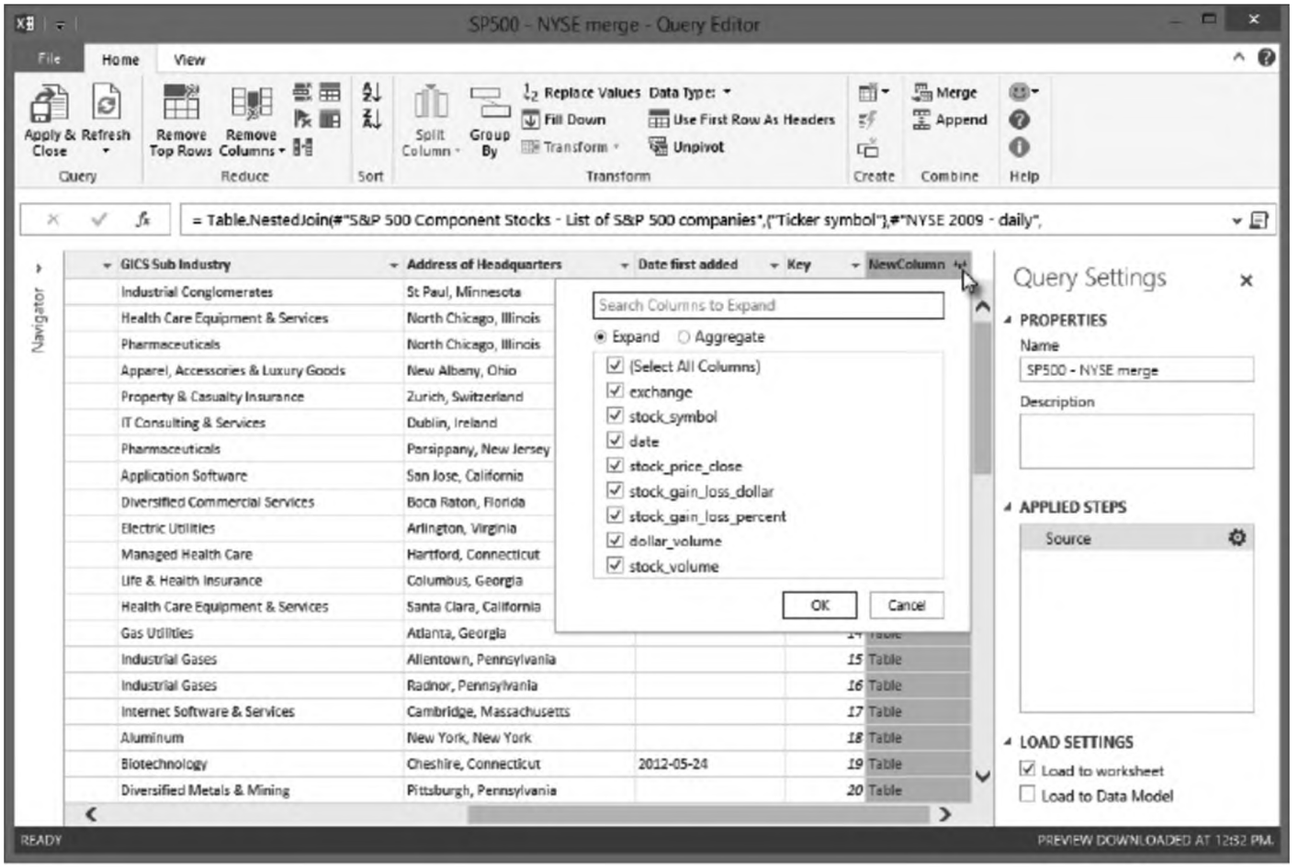


Figura 2-16

Ahora se quiere obtener datos del libro NASDAQ. Todos los datos están en un solo libro de Excel, por lo que es posible obtener datos de un archivo, en lugar de una carpeta. En la cinta de opciones Power Query de Excel, se elige *Obtener datos externos -> Desde archivo -> Desde Excel*. Se selecciona el libro NASDAQ_2009_diario y se hace clic en *Abrir*. En Excel, aparece el panel *Navegador* y al mantener el ratón sobre la hoja de cálculo NASDAQ_2009, se obtiene una vista previa de su contenido (Figura 2-17).

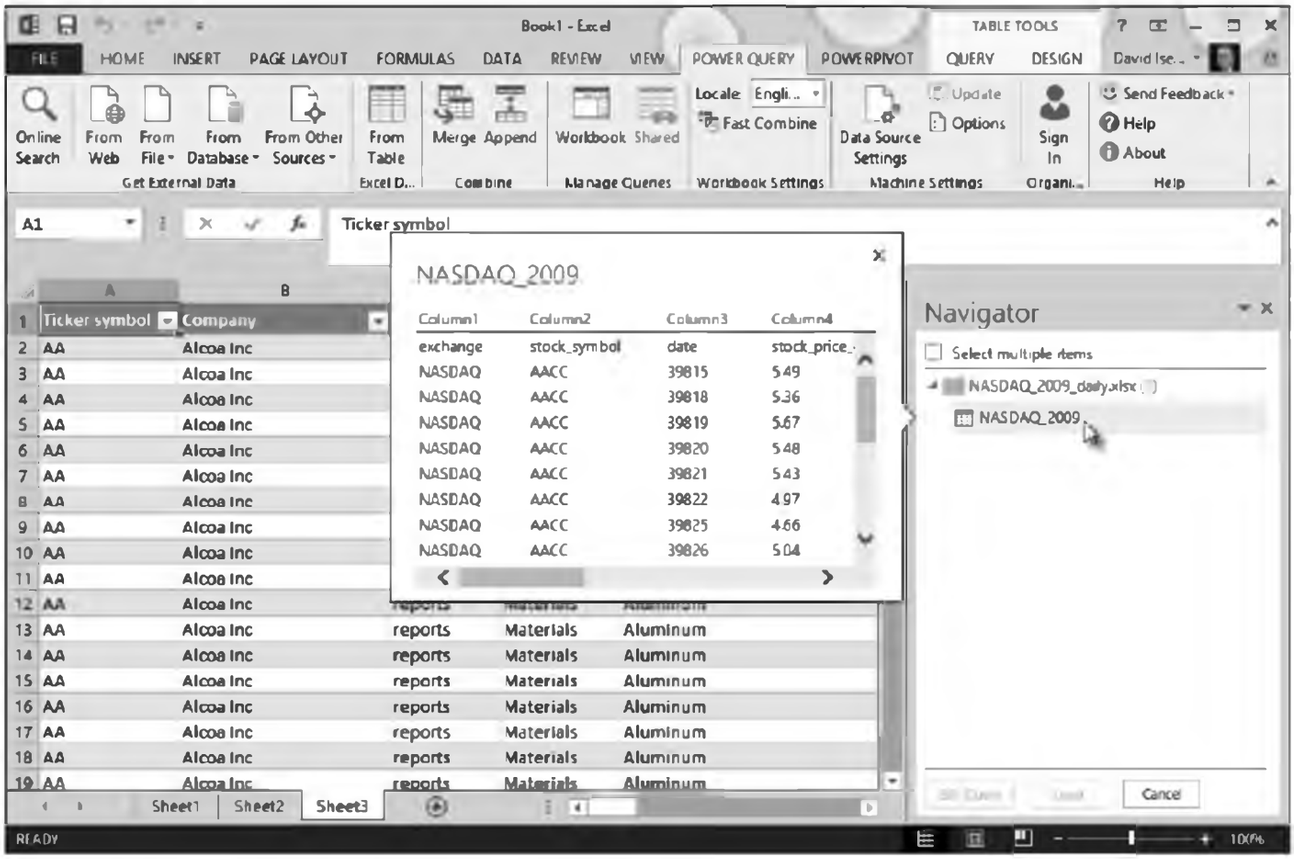


Figura 2-17

Seleccione la hoja de cálculo y haga clic en *Editar consulta* en la parte inferior del panel *Navegador*. Aparece la ventana *Editor de consultas*. De nuevo, desactivamos las casillas de *Configuración de carga* porque quiere dar forma a los datos antes de incluirlos en Excel.

En este punto, los pasos necesarios para mezclar los datos del NASDAQ son similares a los que hizo con los datos del NYSE: active la opción *Usar primera fila como encabezados*, cambie el nombre de la consulta y, a continuación, haga clic en *Aplicar y cerrar*.

Entonces, se tiene una consulta que combina los datos del S&P 500 con los del NYSE y otra consulta que combina los datos del S&P 500 con los del NASDAQ. Pero se quiere que todos esos datos estén en el mismo conjunto de datos. Se necesita anexar esas dos consultas. Para eso, en la barra de herramientas *Consulta de Herramientas de tabla*, se elige *Anexar* (Figura 2-18).

©Alfaomega-RC Libros

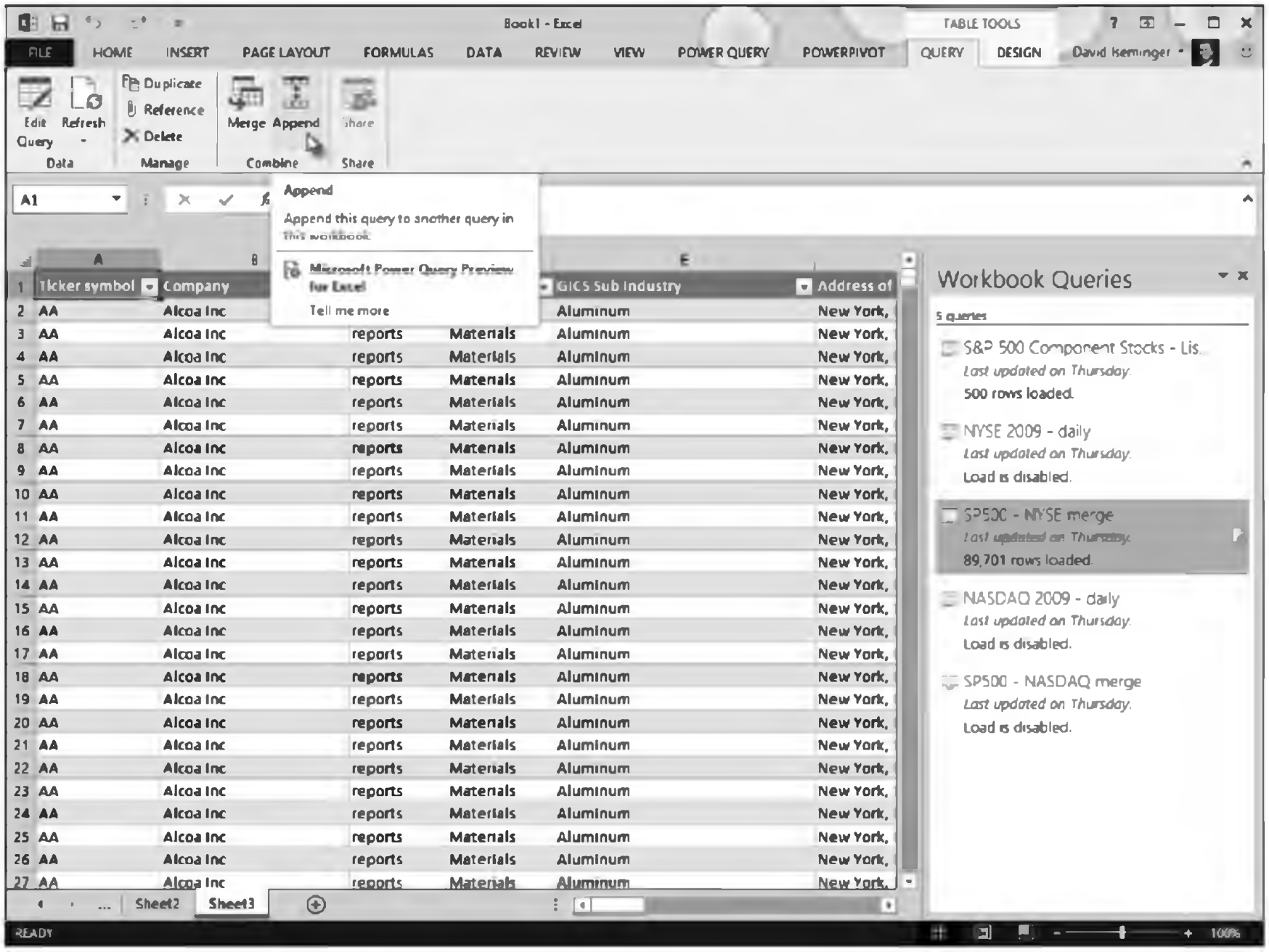


Figura 2-18

Aparece una ventana que le permite seleccionar la tabla principal y la tabla a la que quiere anexar los datos (Figura 2-19). Por suerte, ha ido asignando nombres a las consultas, por lo que le resultará fácil saber qué consultas quiere anexar.

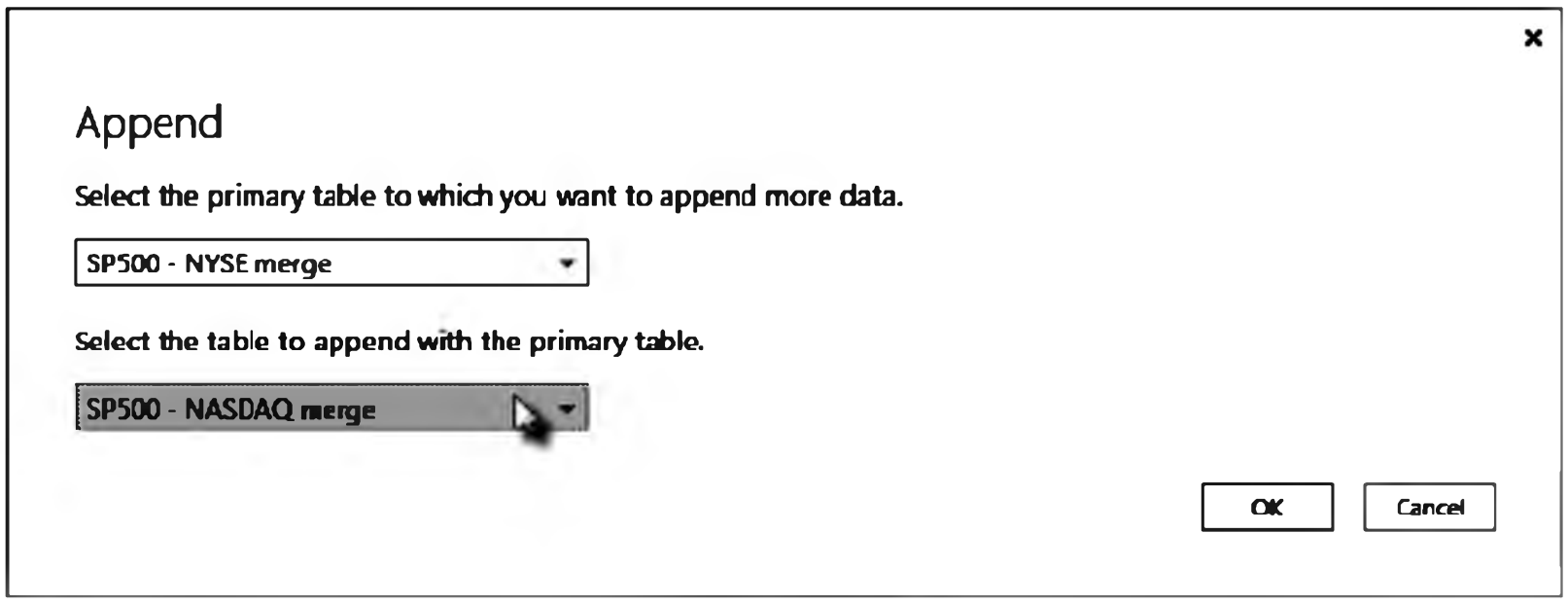


Figura 2-19

Aparece la ventana *Editor de consultas* (Figura 2-20). Se cambia también el nombre de esta consulta y se le llama Anexo NYSE-NASDAQ-SP500 (el título de la ventana *Editor de consultas* se actualiza al escribir el nuevo nombre, y se hace clic fuera del cuadro *Nombre de Configuración de consulta*). El Editor de consultas es la interfaz integral de todas las consultas de Power Query, incluidas las consultas, las combinaciones y los anexos.

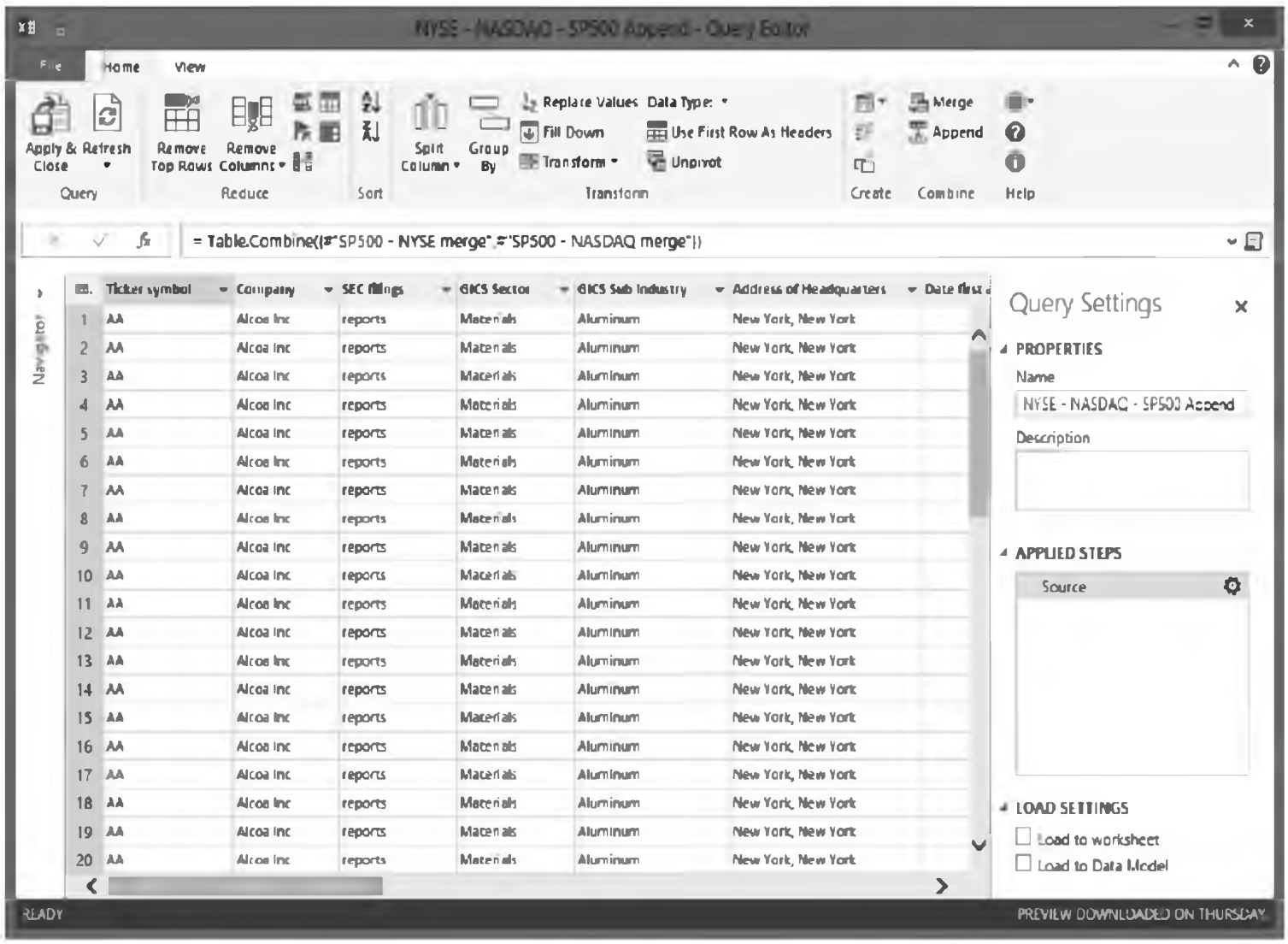


Figura 2-20

Este es el resultado final de los datos que se buscaba: un conjunto de datos diarios del NYSE y el NASDAQ, adaptado para quitar las columnas que no necesita y filtrado para que incluya solo los datos de los símbolos del S&P 500. Ahora se tiene todo lo necesario para cargar los datos en la hoja de cálculo, por lo que se activa la casilla *Cargar en hoja de cálculo* (Figura 2-21). Quedan algunos cambios por realizar antes de agregarlos al modelo de datos, así que deje esa casilla desactivada.

©Alfaomega-RC Libros

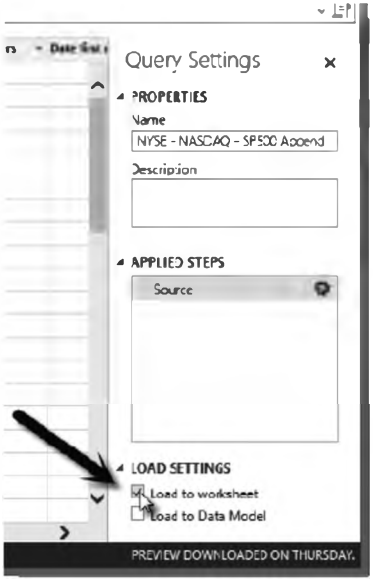


Figura 2-21

Haga clic en *Aplicar y cerrar* en la cinta de opciones *Editor de consultas* y los datos se incorporan a Excel. Con los datos incorporados en Excel, se comprueba que el tipo de datos de todas las columnas es correcto. Por ejemplo, se configura la columna *Fecha* con el tipo de datos *Fecha*. Se cambia también el nombre de algunas columnas y se quita así el prefijo *NuevaColumna*. de cada una de las columnas, para facilitar la lectura (Figura 2-22).

1/2/2009

J	K	L	
date	stock_price_close	stock_gain_loss_dollar	stock_gain_l
1/2/2009	12.11	0	
1/5/2009	11.86	-0.25	
1/6/2009	12.12	0.26	
1/7/2009	10.89	-1.23	
1/8/2009	11.36	0.47	

Figura 2-22

Al finalizar, se elige la pestaña *Power Pivot* de la cinta de opciones y se elige *Tablas -> Agregar a modelo de datos*. Los datos a los que se dio forma, que fueron filtrados y combinados, se cargan en el modelo de datos, quedando listos para poder ser utilizados por otra herramienta de Power BI, por ejemplo Power Pivot.

Hemos visto que con Power Query, se pueden buscar datos en toda la organización y a través de Internet. Después de encontrar los datos que se desean, es posible darles forma, utilizar diferentes conjuntos de datos de distintos orígenes, filtrarlos, combinarlos y anexarlos, todo ello sin tener que incorporar los datos en Excel. Después de darle a la consulta la forma que se quiere y filtrarla, puede descargarse en una hoja de cálculo de Excel, en el modelo de datos o en ambos.

Hay todo tipo de orígenes de datos a los que se puede acceder con Power Query. Se pueden obtener datos externos de bases de datos (Figura 2-23).

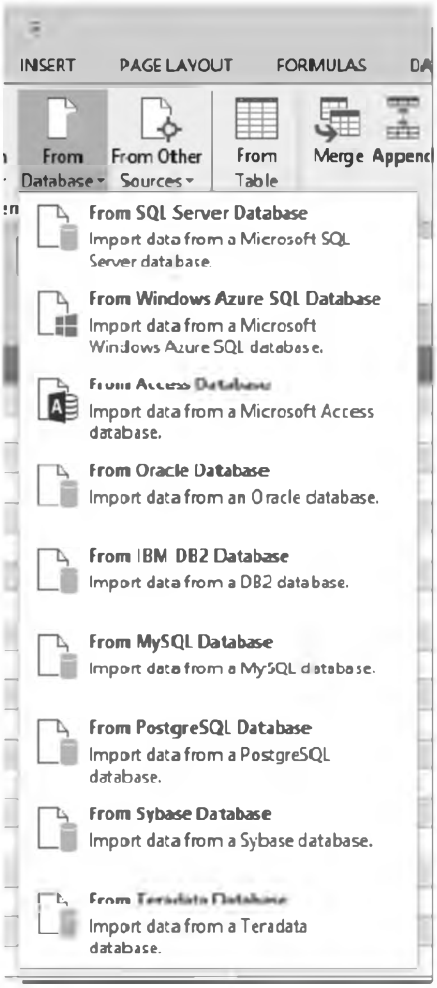


Figura 2-23

También se pueden obtener datos externos de varios archivos, como archivos CSV o archivos de texto (Figura 2-24).

©Alfaomega-RC Libros

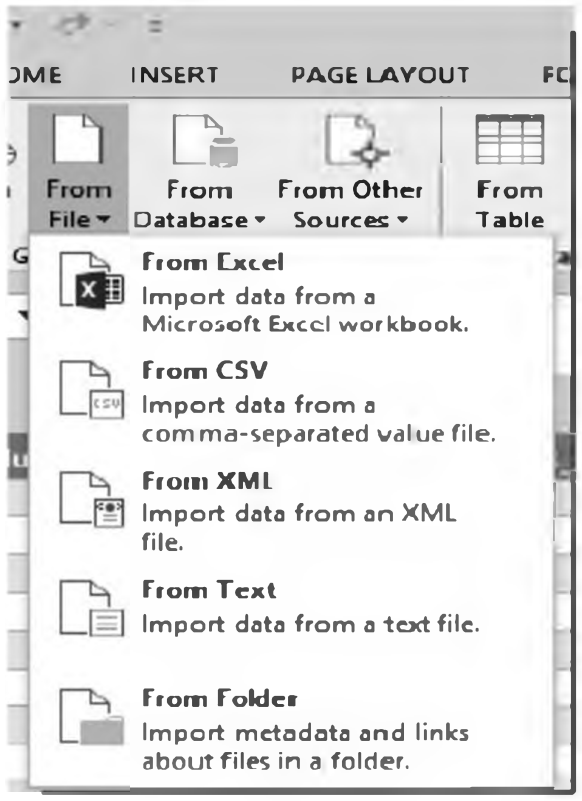


Figura 2-24

Y puede obtener también datos externos de todo tipo de orígenes de datos.

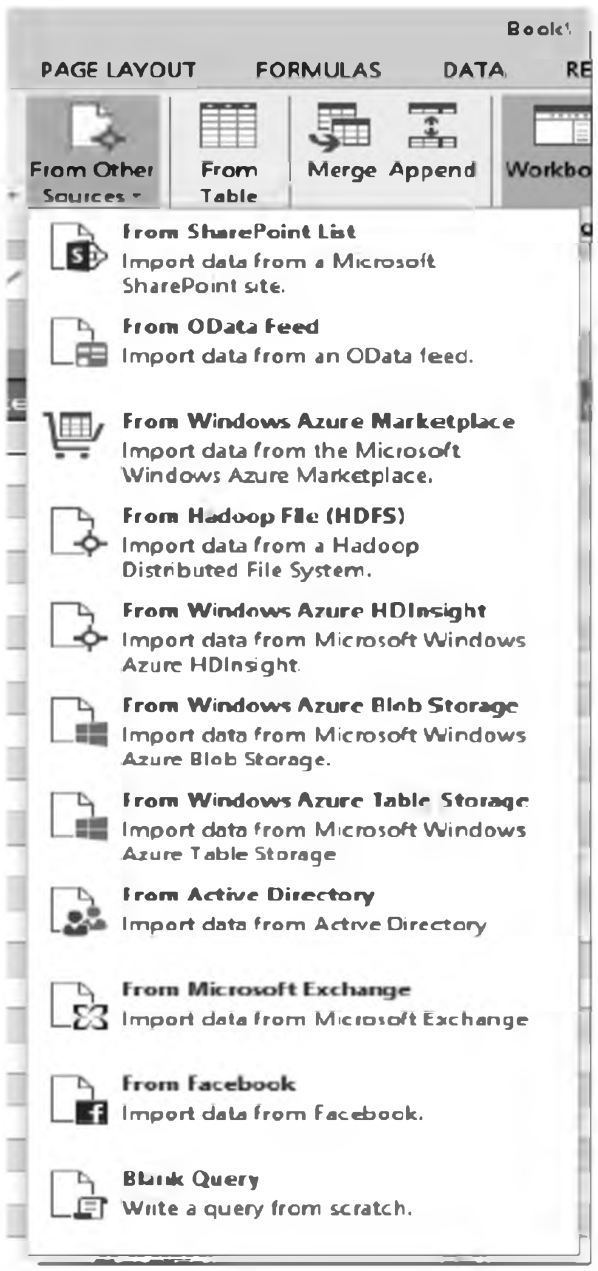


Figura 2-25

Cuando tenga solamente el conjunto de datos que necesita, con la forma que quiere y combinado correctamente, puede guardar la consulta que creó ese conjunto de datos perfeccionado y compartirla con otras personas.

POWER PIVOT

Con Power Pivot, puede crear su propio modelo de datos a partir de varios orígenes de datos, modelados y estructurados con precisión para que se ajusten a sus necesidades, y puede actualizarlos desde sus fuentes originales siempre que quiera. Ahora, ya tiene los datos cargados en Power Pivot, la eficaz función de modelado de datos con características de autoservicio de Excel. Power Pivot permite crear y administrar una colección de tablas y relaciones, desde dentro de Excel.

Podemos cargar los datos diarios anexados del S&P 500 y el NASDAQ en el modelo de datos, pero también queremos cargar la tabla del S&P 500 de base. Para ello se elige la pestaña de la hoja de cálculo de Excel en la que cargó la lista del S&P 500 y, luego, en la pestaña *Power Pivot* de la barra de herramientas, se elige *Agregar a modelo de datos* (Figura 2-26).

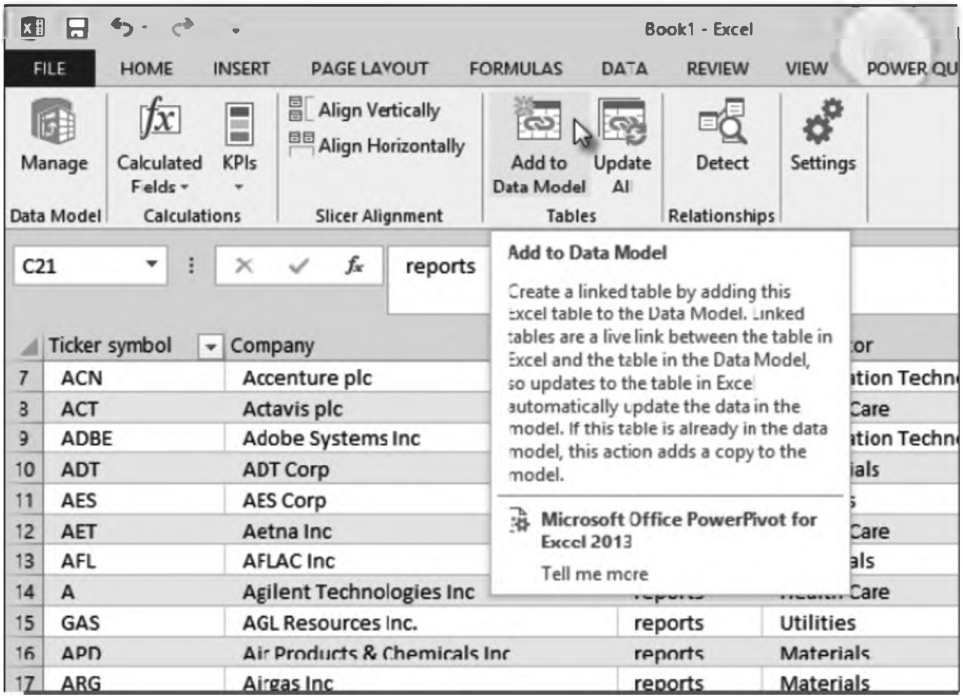


Figura 2-26

©Alfaomega-RC Libros

Cuando se agrega la tabla al modelo de datos, Power Pivot se abre en una ventana independiente y muestra la tabla que agregó al modelo. También puede abrir la ventana Power Pivot eligiendo el icono *Administrar* de la barra de herramientas *Power Pivot* (Figura 2-27).

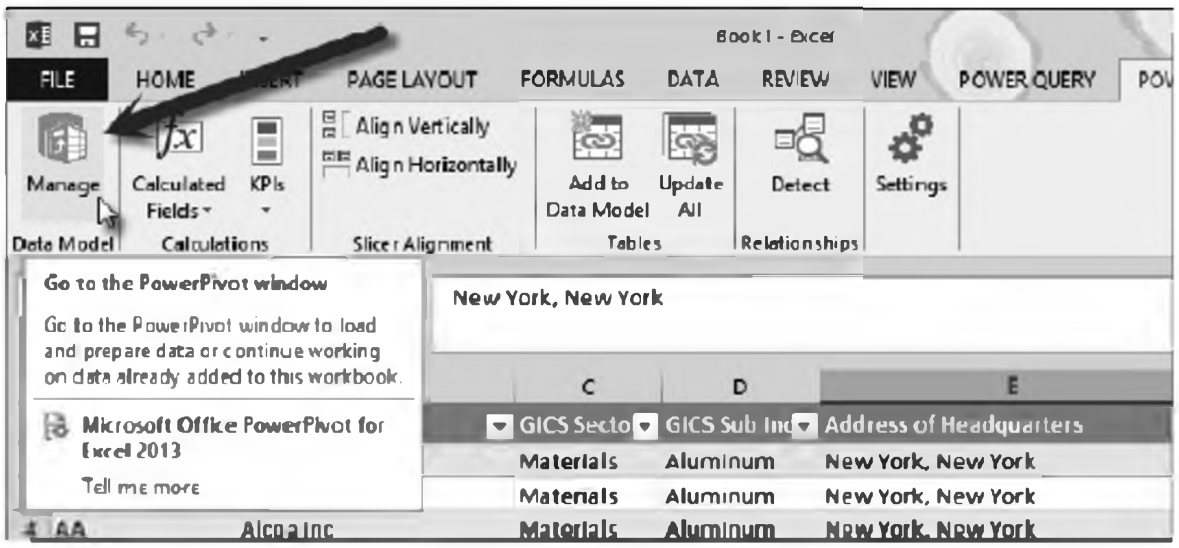


Figura 2-27

En Power Pivot, las tablas del modelo de datos aparecen en pestañas, de forma similar a como aparecen en Excel. Ahora se decide cambiar el nombre de la tabla que acaba de agregar a SP 500, por lo que se hace doble clic en esa pestaña y se escribe el nuevo nombre (Figura 2-28).

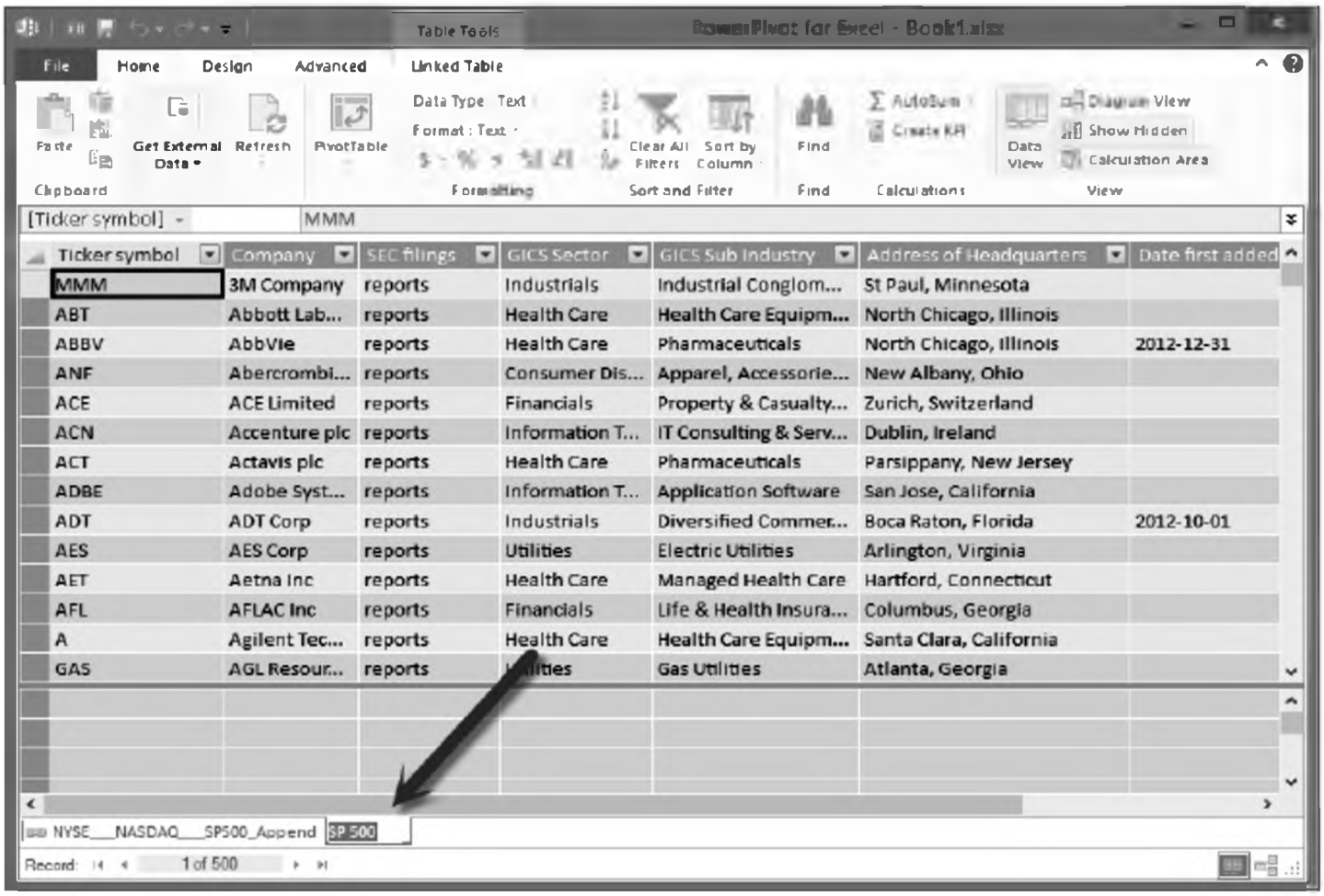


Figura 2-28

Mientras se revisan los datos del modelo, se decide que falta una perspectiva importante: el rendimiento anual. Podrían usarse los datos del primer día, como el precio, por ejemplo, y los del último día y, simplemente, observar la diferencia. Eso sería bastante fácil y rápido.

Ya se tiene una consulta que combina y anexa todos los datos diarios, que es la consulta que se llamó Anexo NYSE-NASDAQ-SP500. Se necesita una versión de esa misma consulta que filtre aún más los datos, para que solo aparezcan el primer día y el último. Se vuelve a Excel y se busca la consulta. Para cambiar entre Power Pivot y Excel basta con elegir la ventana Excel. No es necesario cerrar la ventana Power Pivot.

De nuevo en Excel, se ve que se cerró el panel *Consultas del libro*. Para recuperar ese panel, se elige *Power Query -> Administrar consultas -> Libro*. Se observa que, si se vuelve a elegir el botón de la cinta Libro, se cierra el panel *Consultas del libro*, pero se quiere que quede visible, así que se hace clic otra vez en el botón *Libro* (Figura 2-29).

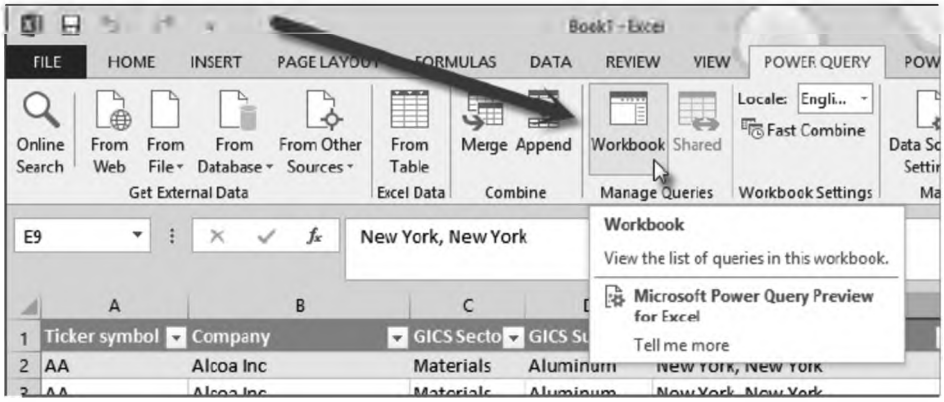


Figura 2-29

Una vez hallada la consulta en el panel *Consultas del libro* , se va a la pestaña *Consulta de Herramientas de tabla* y se elige *Duplicar* (Figura 2-30). Así, se puede empezar por la consulta que se creó para anexar los datos combinados del NYSE y el NASDAQ y, después, filtrarlos para que incluyan solo el primer día de cotización del año y el último. La consulta duplicada aparece en la parte inferior del panel *Consultas del libro* (Figura 2-31), como Anexo NYSE-NASDAQ-SP500 (2). Cuando se mantiene el ratón sobre esa consulta, aparece un recuadro con una vista previa de los datos. Elija *Editar consulta* en la parte inferior del recuadro.

©Alfaomega-RC Libros

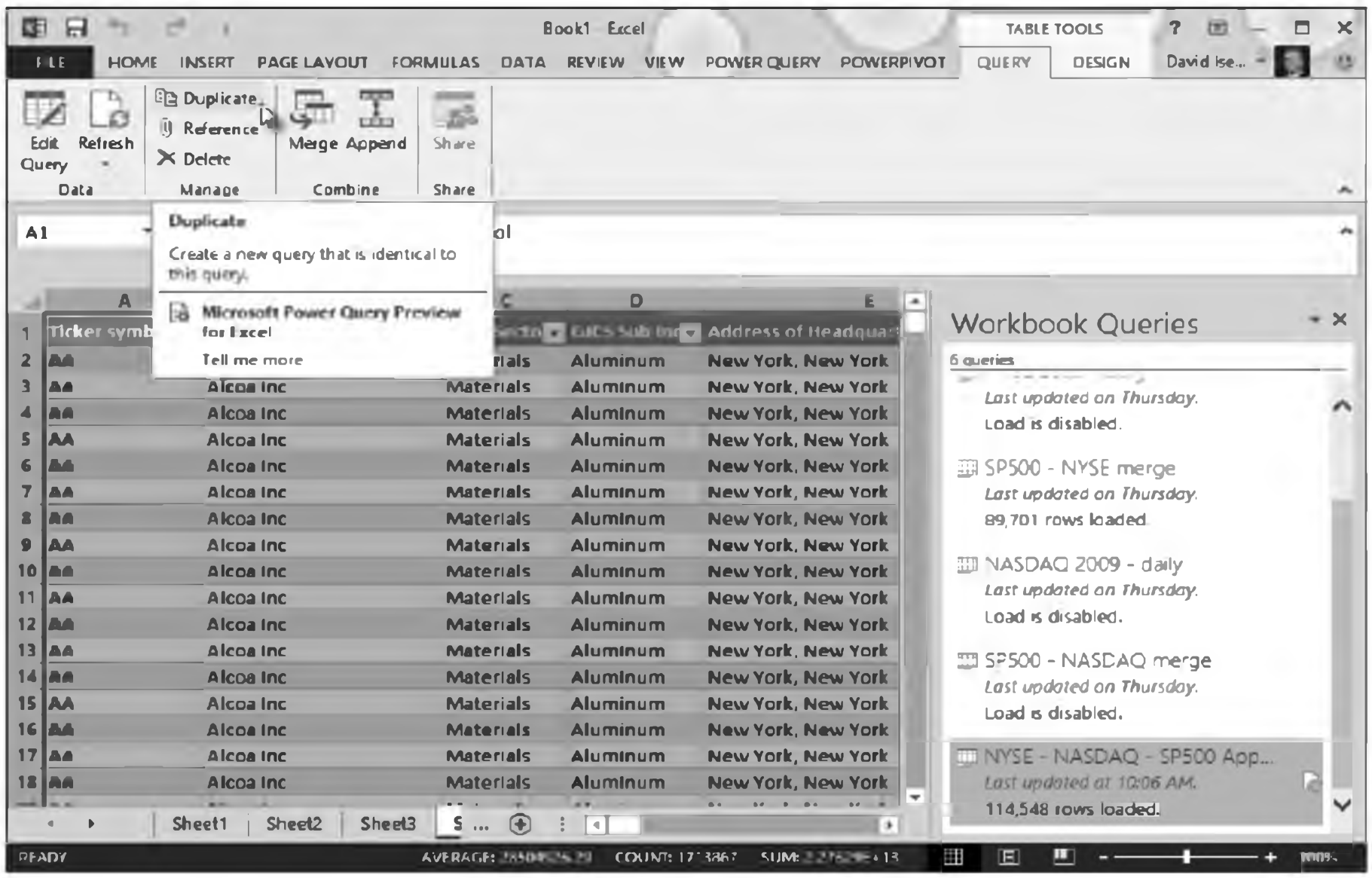


Figura 2-30

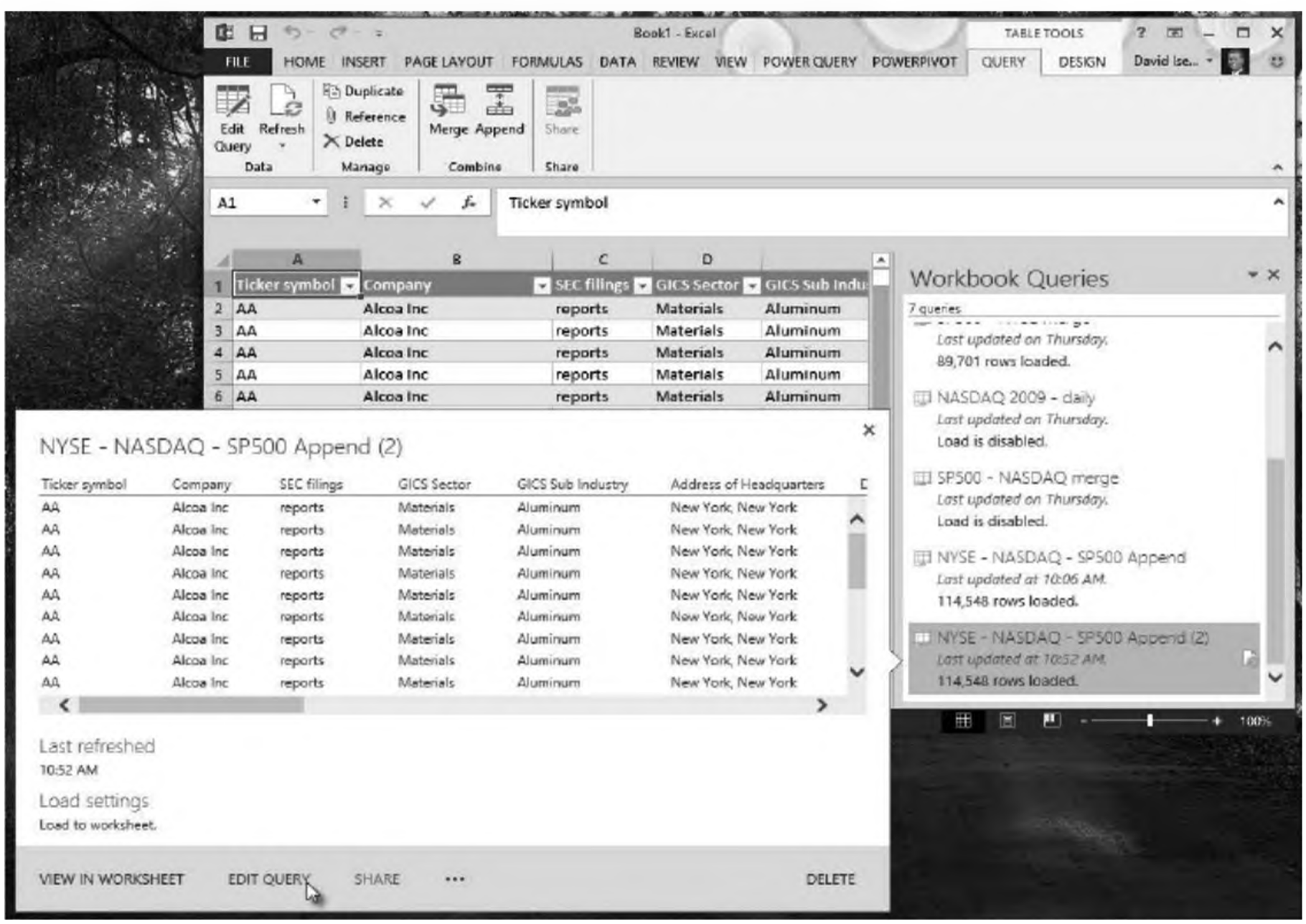


Figura 2-31

Aparece la ventana *Consulta* y, desde ahí, se puede restringir el campo NuevaColumna.fecha de modo que incluya solo el primer día de cotización. Se va al campo NuevaColumna.fecha y se observa que el tipo de datos de ese campo no tiene el formato *Fecha* (Figura 2-32).

Es fácil corregirlo, y se puede hacer directamente en el Editor de consultas. Para ello se elige la columna NuevaColumna.fecha y, a continuación, se selecciona *Fecha* en la lista desplegable *Tipo de datos*: de la sección *Transformar* de la cinta de opciones (Figura 2-33).

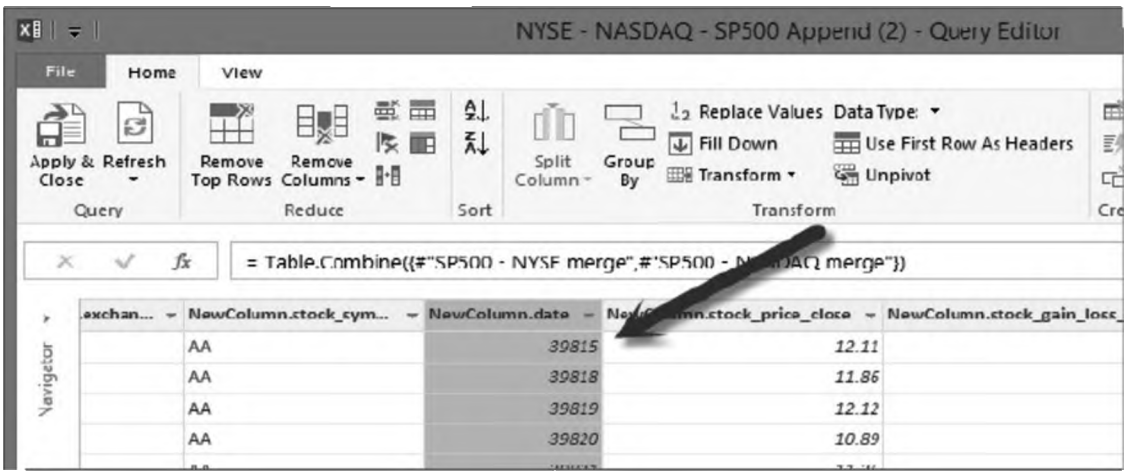


Figura 2-32

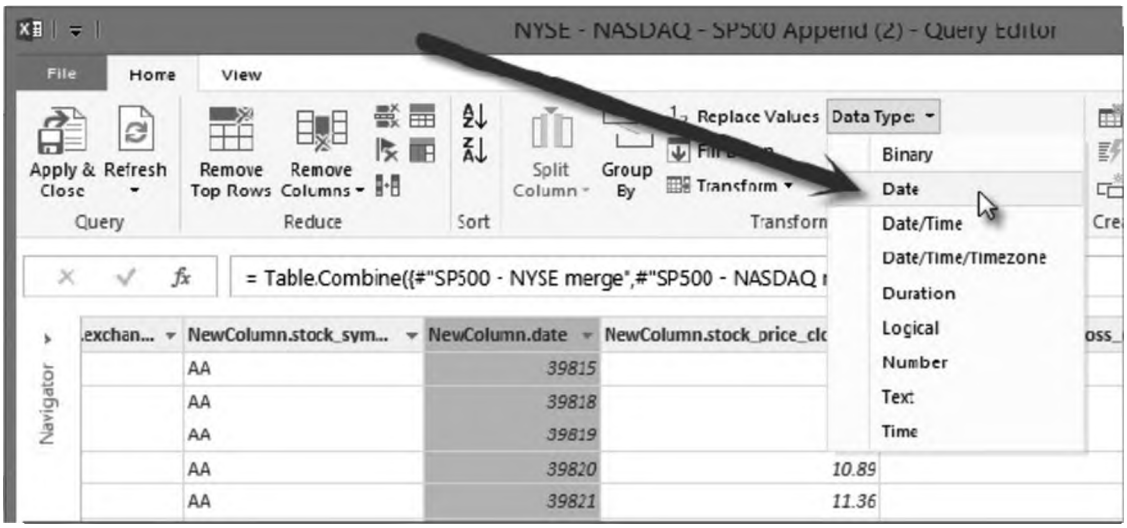


Figura 2-33

Con el tipo de datos configurado como Fecha, se puede filtrar la fecha de esa columna para que contenga solo el primer día de cotización del año (Figura 2-34).

©Alfaomega-RC Libros

Haga clic con el botón secundario en la consulta Primer día de cotización en el panel *Consultas del libro* y elija *Duplicar*. La consulta duplicada se agrega al panel Consultas del libro y se titula Primer día de cotización (2). Es posible que tenga que desplazarse hacia abajo por el panel Consultas del libro para verla. Haga clic con el botón secundario del ratón en *Primer día de cotización* (2) y elija *Editar consulta* en el menú que aparece.

Los pasos siguientes son un poco más complicados, pero demuestran la flexibilidad y la amplia funcionalidad de Power Query. Además, muestran una característica muy útil de Power Query: cómo administra los pasos de dar forma a la consulta y filtrarla.

Cuando se elige *Editar consulta* en Primer día de cotización (2), aparece la siguiente pantalla (Figura 2-35). Observe el campo *Fecha* y la sección *Pasos aplicados* del panel *Configuración de consulta*, donde está seleccionada la última entrada (ColumnasQuitadas).

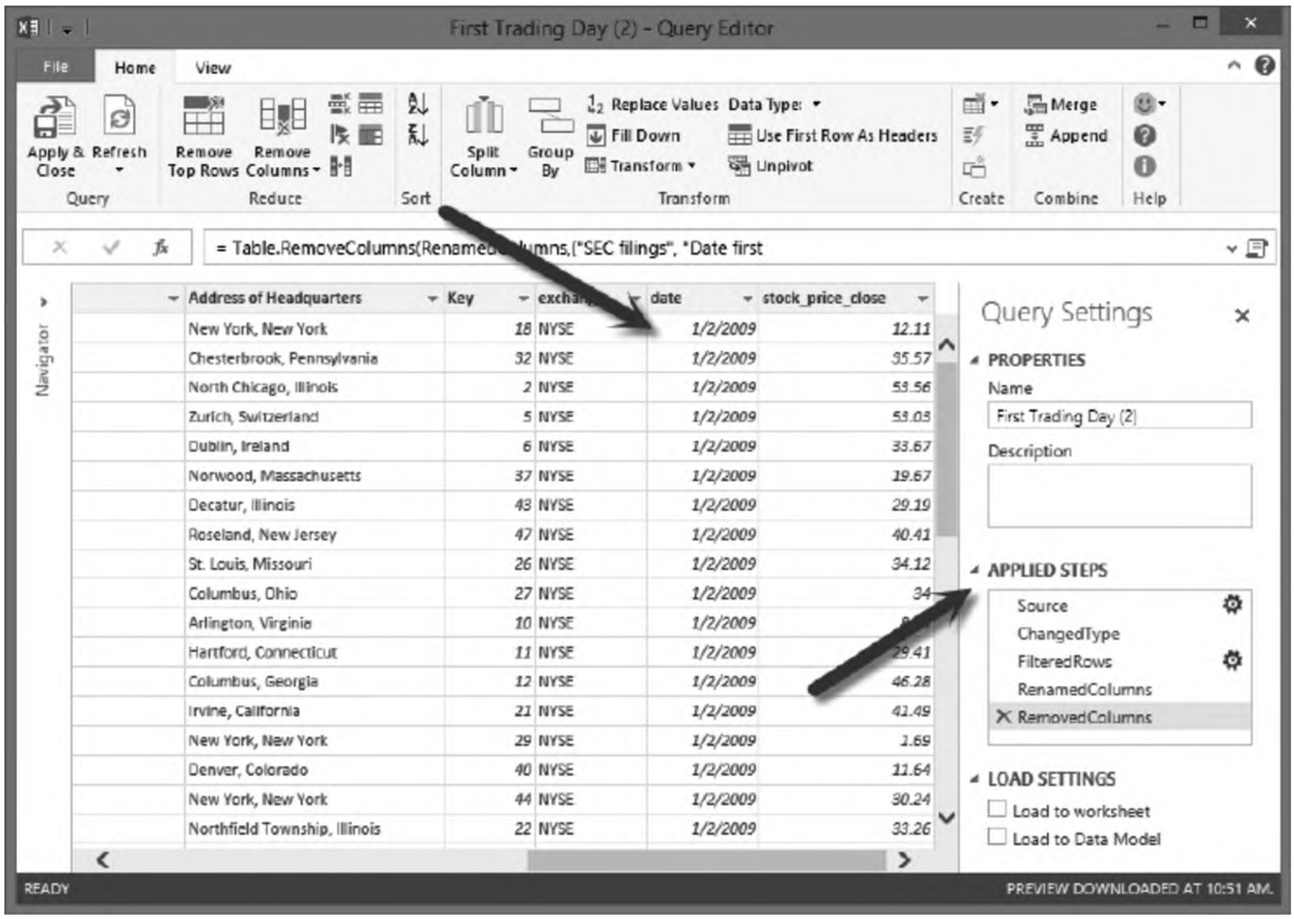


Figura 2-35

©Alfaomega-RC Libros

Cuando se elige uno de los pasos anteriores, los datos de la consulta recuperan la forma (y la ausencia de filtros aplicados) que tenían en ese punto del proceso de modelado de la consulta. Por ejemplo, cuando elige *Origen* (el primer paso aplicado de Pasos aplicados), los datos vuelven al estado en el que estaban cuando se aplicó ese paso a la consulta (Figura 2-36).

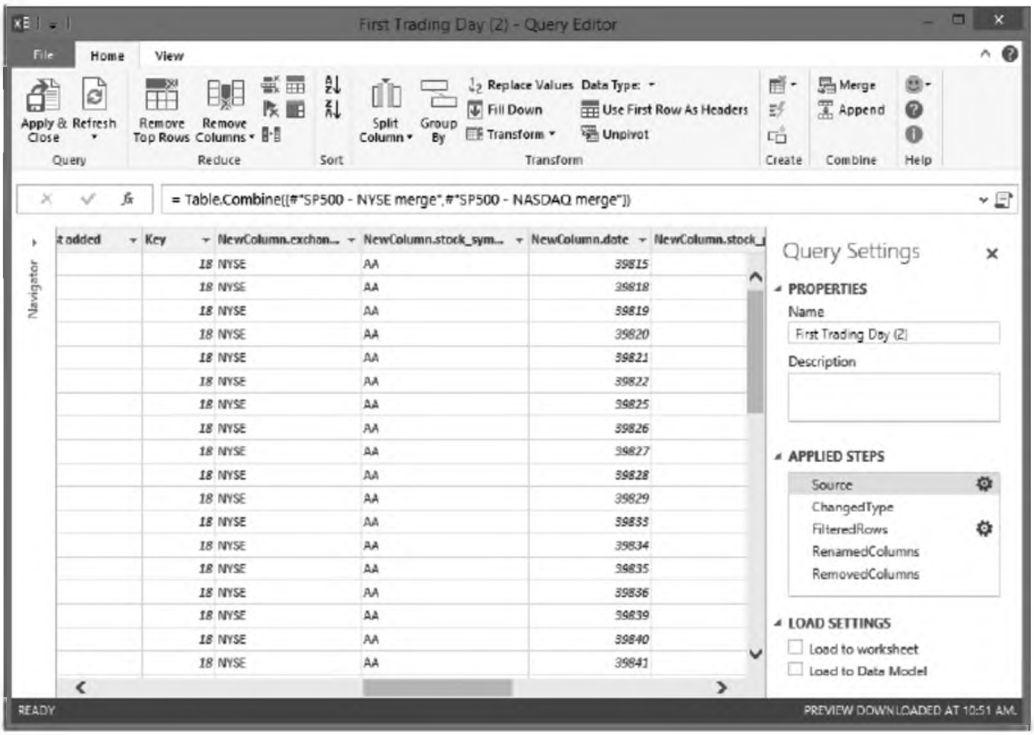


Figura 2-36

Observe que los nombres de las columnas volvieron a su estado anterior, y ahora tienen el prefijo NuevaColumna otra vez. Observe también que el campo de fecha recuperó el tipo de datos general y (aunque es difícil saberlo) no hay ningún filtro aplicado al campo de fecha, de modo que todos los días de cotización se encuentran actualmente en el conjunto de datos. Cuando elige el cuarto paso de Pasos aplicados (el paso NombreDeColumnasCambiado), los datos aparecen con la forma que se les aplicó en ese paso del proceso de dar forma a los datos (Figura 2-37). En ese momento, los únicos cambios restantes son los de quitar las columnas que no se usan.

Ahora se quiere cambiar la consulta de forma que filtre los resultados para que aparezcan solo los del último día de cotizaciones del año, en lugar del primer día del año. Se tienen en cuenta los siguientes pasos:

Haga clic en el icono del engranaje, a la derecha de la entrada FilasFiltradas, en la sección Pasos aplicados del panel *Configuración de consulta*. Aparece la ventana de la Figura 2-38. Elija la flecha desplegable del campo que muestra 2/1/2009 y seleccione la última entrada, que es 31/12/2009. Seleccione *Aceptar*. Elija el último elemento de Pasos aplicados (el paso de ColumnasQuitadas). Cambie el nombre de la consulta a Último día de cotización. Elija *Aplicar y cerrar*.

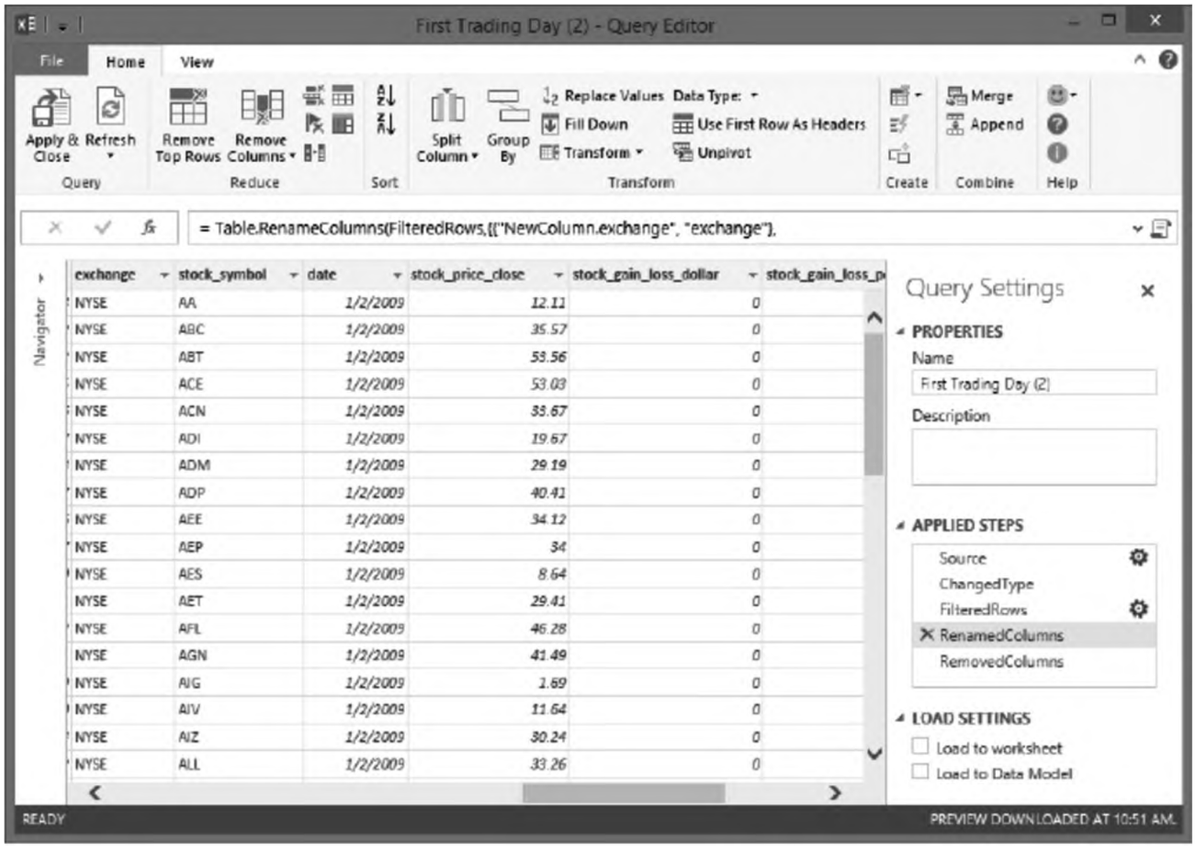


Figura 2-37

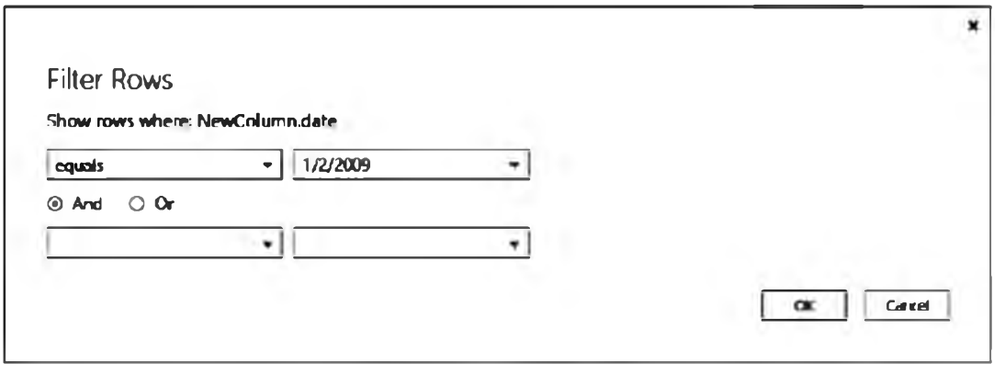


Figura 2-38

A continuación, es necesario combinar estas dos consultas. Estos son los pasos que se siguen:

©Alfaomega-RC Libros

En la pestaña *Power Query* de la cinta de opciones, elige *Combinar* (Figura 2-39). Aparece la ventana *Combinar*. Especifique las consultas Primer día de cotización y Último día de cotización como las tablas para combinar y, luego, elija Símbolo del valor como columnas coincidentes (Figura 2-40). Elija *Aceptar*.

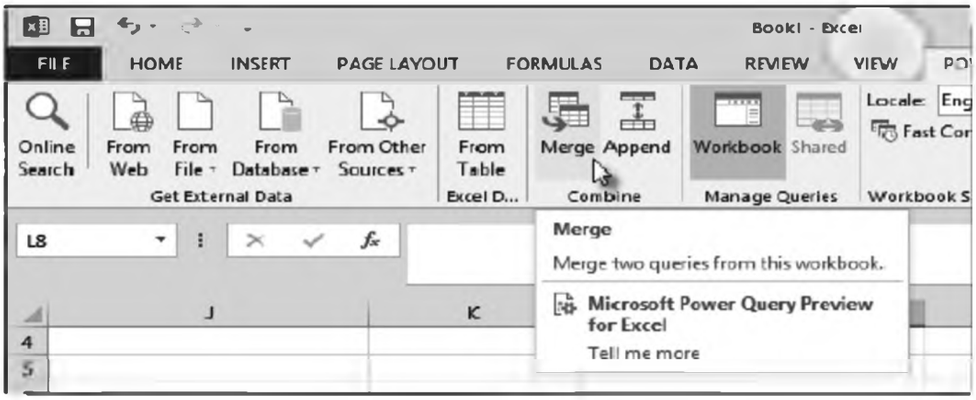


Figura 2-39

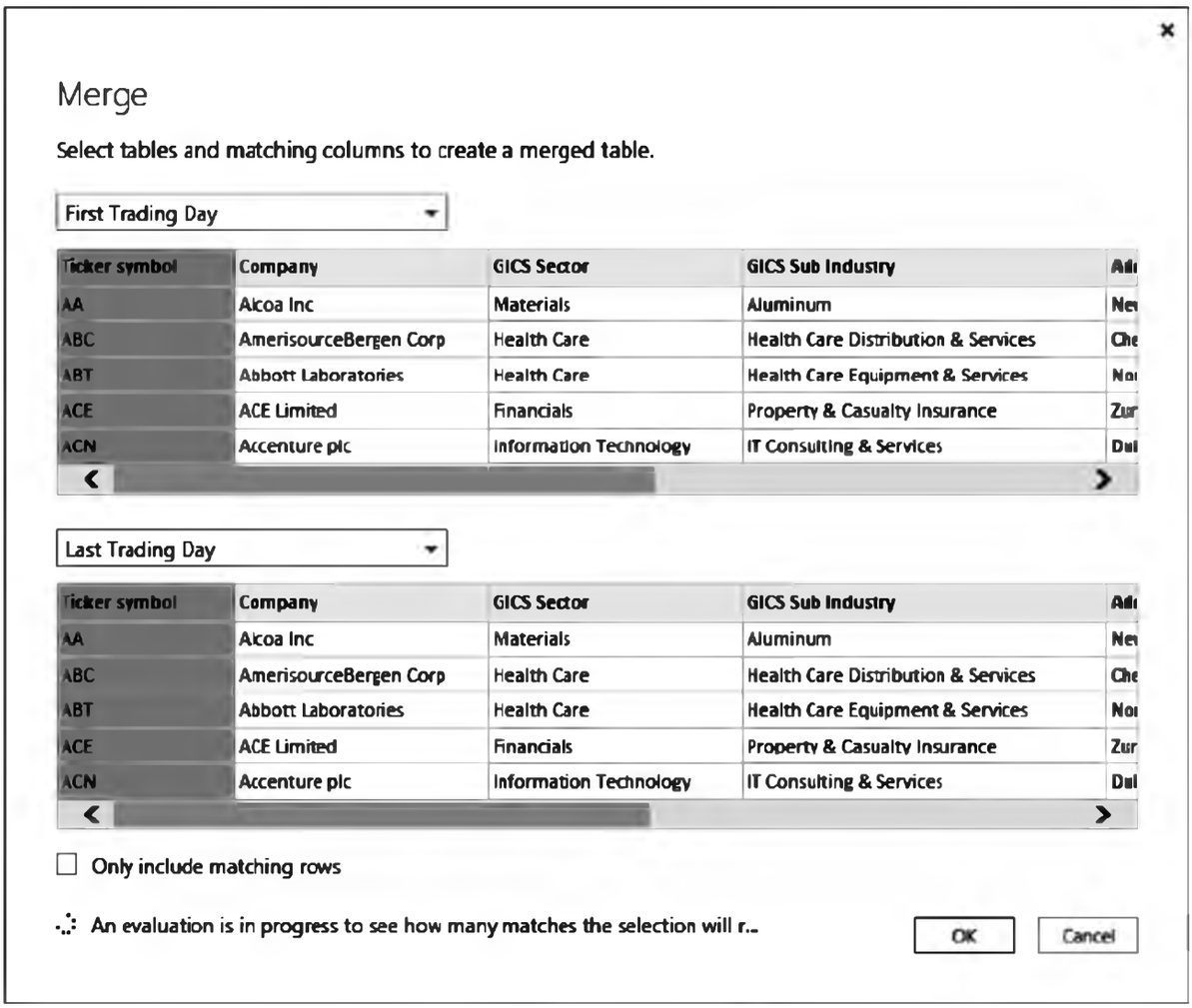


Figura 2-40

Como antes, las columnas de la consulta combinada aparecen como una tabla, bajo una columna titulada NuevaColumna que se encuentra al final de la consulta. Haga clic en el icono de flecha doble hacia fuera para expandir esa tabla en columnas individuales (Figura 2-41).

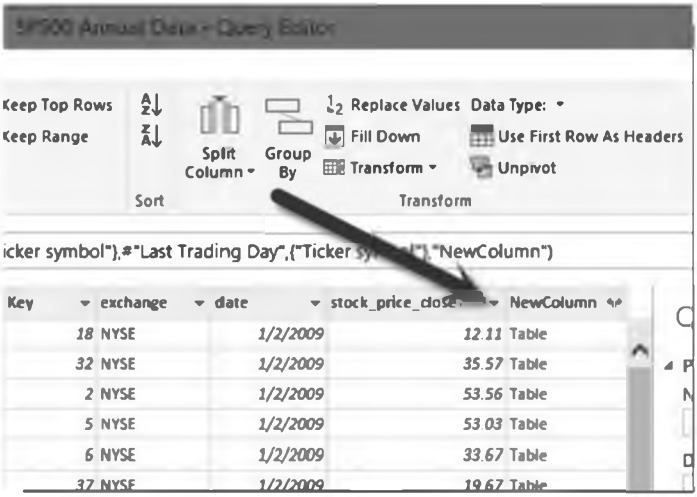


Figura 2-41

Cambie el nombre de la consulta a Datos anuales de SP500. Active la casilla que hay junto a *Cargar en modelo de datos* y, a continuación, elija *Aplicar y cerrar*.

Cuando se termina, se tiene una nueva tabla en el modelo de datos, que contiene los datos del precio de cierre del primer día y el último. La tabla del modelo de datos contiene algunas columnas que no necesita y otras columnas que quiere cambiar de nombre. Se puede hacer todo esto desde dentro de Power Pivot.

En Power Pivot, haga los cambios siguientes:

- Cambie el nombre de la columna fecha a Fecha del primer día y el de la columna NuevaColumna.fecha a Fecha del último día.
- Cambie el nombre de precio_cierre_valor a Cierre del primer día y el de NuevaColumna.precio_cierre_valor a Cierre del último día.
- Quite las siglas GICS de los nombres de columna Sector GICS y Subsector GICS.
- Elimine las columnas siguientes, ya que son redundantes (también podría haber hecho esto en Power Query): Clave, NuevaColumna.Símbolo de bolsa,

 ©Alfaomega-RC Libros

NuevaColumna.Empresa, NuevaColumna.Sector GICS, NuevaColumna.Subsector GICS, NuevaColumna.Dirección de la sede, NuevaColumna.Clave y NuevaColumna.bolsa.

Cambiar el nombre de estas columnas (Figura 2-42) es conveniente para crear infomes a partir de datos.

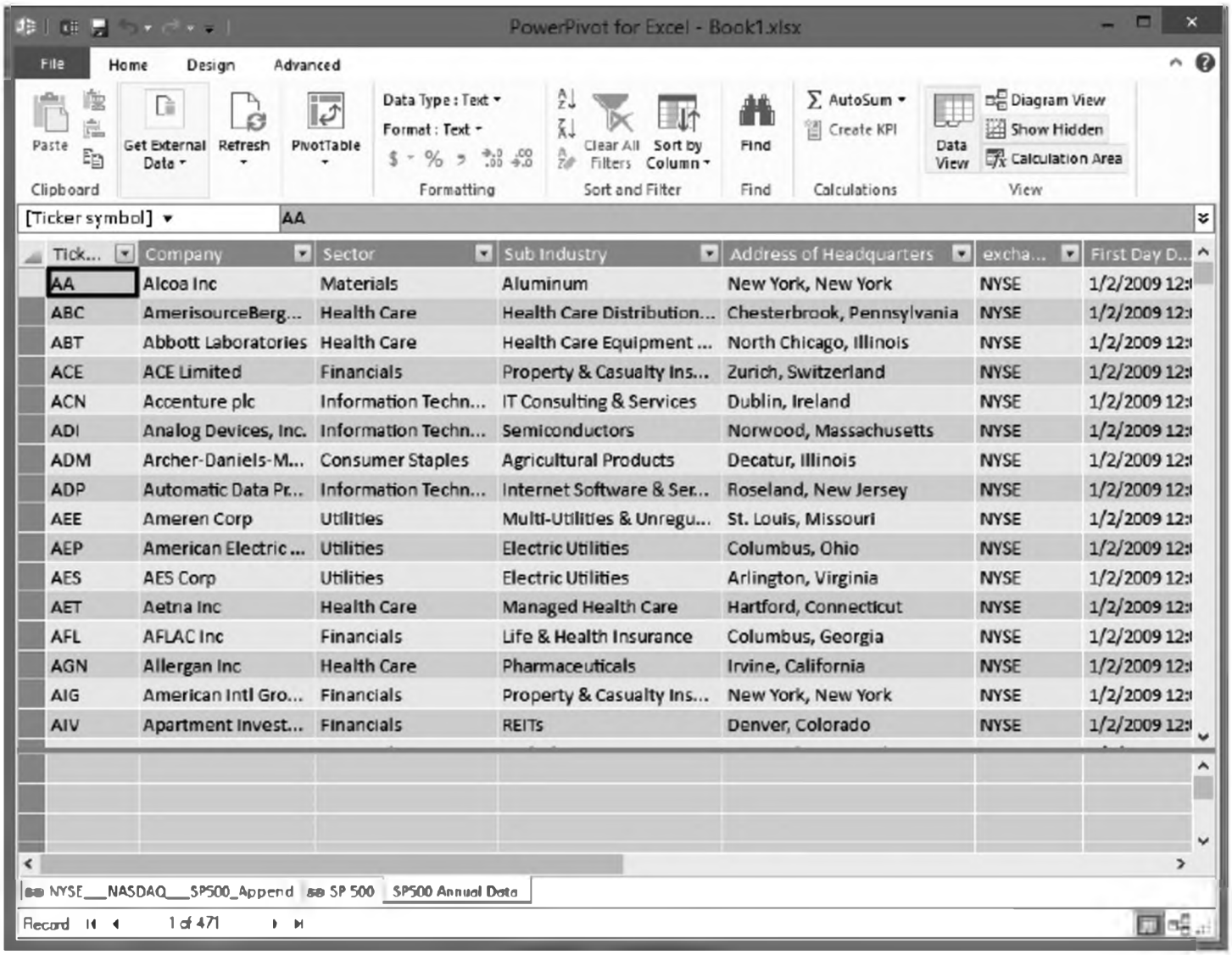


Figura 2-42

Crear cálculos

Una vez que se tienen los datos anuales en el modelo de datos, se puede calcular la ganancia global, tanto en dólares como en porcentaje, de cada valor. En Power Pivot, puede crear columnas nuevas en cualquier tabla, y usar en esas columnas fórmulas que realizan muchas funciones diferentes. Estos cálculos se suelen denominar “campos calculados”.

En Power Pivot, los campos calculados usan fórmulas de Expresiones de análisis de datos (DAX). Las fórmulas de DAX son muy similares a las fórmulas de Excel y funcionan en Power Pivot y en las tablas dinámicas. Si ya está familiarizado con la creación de fórmulas de Excel, con DAX se sentirá como en casa.

La primera fórmula que se crea calcula la ganancia anual, o la pérdida, en dólares. Se comienza por seleccionar una celda de *Agregar columna* y, después, se escribe la fórmula en la barra de fórmulas de DAX. A medida que se escribe, Power Pivot ofrece sugerencias basadas en las tablas disponibles y los campos del modelo de datos (Figura 2-43).

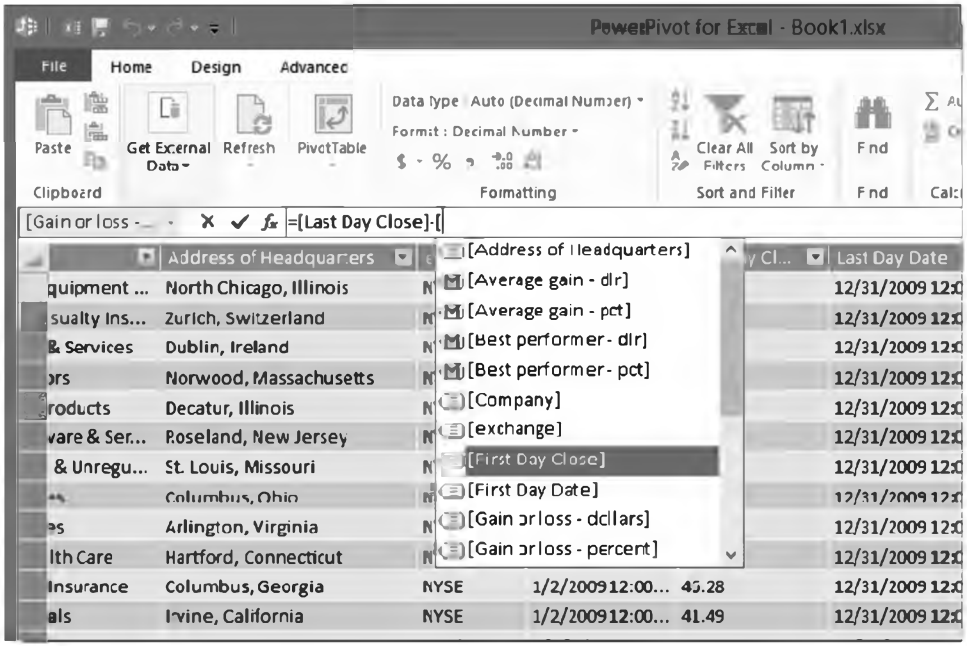


Figura 2-43

Al completar la fórmula, se calculan los resultados de cada fila de la tabla. Se agregan algunas columnas más, como la ganancia anual en porcentaje, y se aplica el tipo de datos y el formato adecuados para cada una de ellas. Estos son los cálculos que se crean (con sus fórmulas, entre paréntesis a continuación):

- Ganancia o pérdida en dólares (=[Cierre del último día]-[Cierre del primer día])
- Ganancia o pérdida en porcentaje (=([Cierre del último día]-[Cierre del primer día])/[Cierre del primer día])

©Alfaomega-RC Libros

- Valores con mejores y peores resultados (=RANKX('Datos anuales de SP500',[Ganancia o pérdida en dólares],))

Además de las nuevas columnas, se quieren crear algunos campos calculados que proporcionen sumas de las cantidades, por ejemplo. Esto se hace seleccionando una celda del área de cálculo de Power Pivot, que es la colección de celdas que hay entre los datos de la tabla y las pestañas. Los campos calculados pueden ir en cualquier celda del área de cálculo (Figura 2-44).

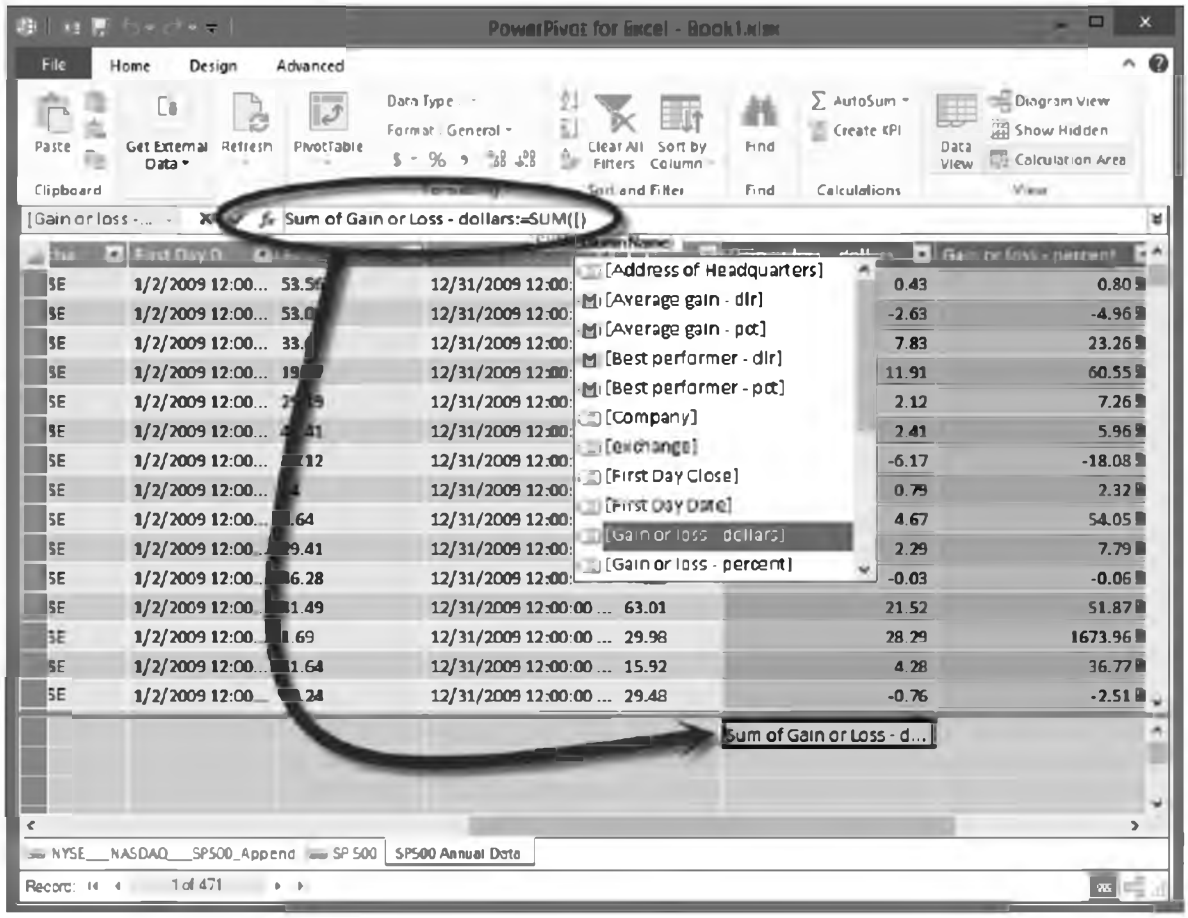


Figura 2-44

Se crean los siguientes campos calculados:

- Suma de ganancia o pérdida en dólares: =SUMA([Ganancia o pérdida en dólares])
- Ganancia promedio en dólares: =PROMEDIO([Ganancia o pérdida en dólares])
- Ganancia promedio en porcentaje: =PROMEDIO([Ganancia o pérdida en porcentaje])
- Valor con mejores resultados en dólares: =MAX([Ganancia o pérdida en dólares])
- Valor con peores resultados en dólares: =MIN([Ganancia o pérdida en dólares])

- Valor con mejores resultados en porcentaje: =MAX([Ganancia o pérdida en porcentaje])
- Valor con peores resultados en porcentaje: =MIN([Ganancia o pérdida en porcentaje])

Crear relaciones

En Power Pivot, también puede definir relaciones entre las tablas. Las relaciones establecen una conexión entre las tablas, basada en una columna de cada una de las tablas que tiene datos idénticos o similares. Las relaciones permiten crear informes que incluyen datos de tablas relacionadas.

A continuación se crea una relación entre la tabla SP 500 y cada una de las otras dos tablas que tiene en el modelo de datos, a partir del símbolo del valor. Puede hacer esto con solo arrastrar y colocar, en la Vista Diagrama de Power Pivot. En Power Pivot, elija *Vista Diagrama* en la sección *Vista* de la cinta de opciones Inicio (Figura 2-45).

Figura 2-45

Para crear una relación, arrastre el campo de la tabla principal al campo correspondiente de la tabla con la que se debe establecer la relación. Una línea conecta las tablas para indicar la relación. Puede seleccionar la línea para resaltar los campos de la relación.

Creamos una relación arrastrando el campo Símbolo del valor de la tabla Datos anuales de SP500 hasta el campo Símbolo del valor de la tabla SP500. Observando el nombre de la otra tabla en la Vista Diagrama, se ve que es bastante largo: Anexo_NYSE___NASDAQ___SP500. Cambiamos el nombre en la propia Vista Diagrama. Para ello, se hace doble clic en el nombre de la tabla y se escribe el nuevo nombre: Datos diarios de SP500.

©Alfaomega-RC Libros

Arrastre el campo Símbolo del valor de Datos diarios de SP500 hasta el Símbolo del valor de la tabla SP 500 y cree otra relación. Al seleccionar la línea de relación que une las tablas, se resalta la relación. Ahora, la Vista Diagrama de Power Pivot aparece como la Figura 2-46.

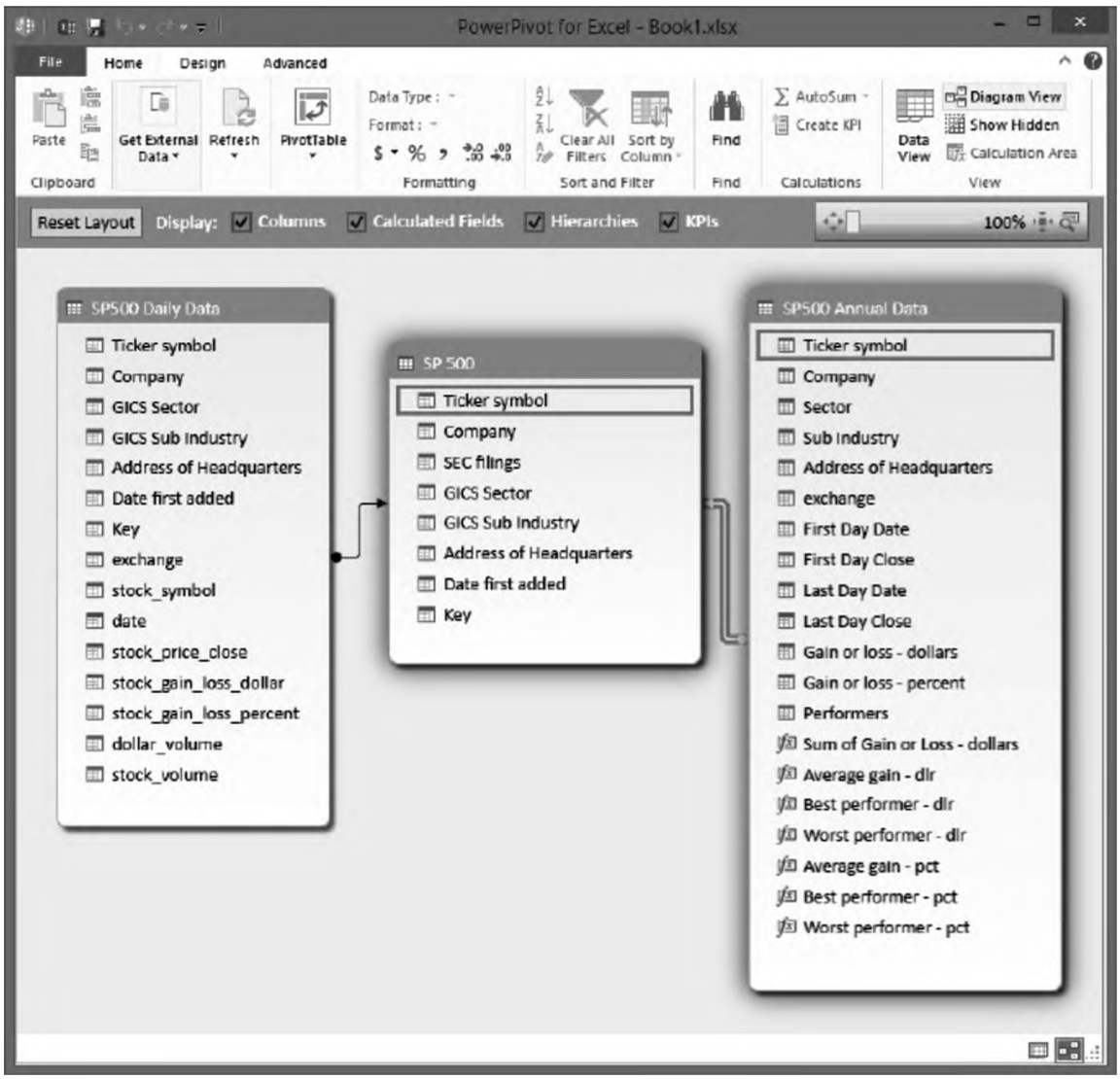


Figura 2-46

Crear jerarquías

En Power Pivot, las jerarquías son agrupaciones de elementos de datos que comparten una relación lógica vertical. Por ejemplo, una jerarquía geográfica podría ser comunidad autónoma, provincia y localidad; la comunidad autónoma está por encima de la provincia en la jerarquía (y contiene una o varias provincias), y la provincia se encuentra por encima de la localidad (y contiene muchas localidades).

Usar jerarquías en Power Pivot permite crear informes para explorar los datos en profundidad. Consideramos que tratar el sector y el subsector como una jerarquía sería interesante, especialmente si se puede crear un informe que permita a los usuarios explorar en profundidad un sector determinado. Hay varias maneras de crear jerarquías en Power Pivot. Decidimos crear una jerarquía en la Vista Diagrama.

Para crear una jerarquía en la tabla Datos anuales de SP500, hacemos clic con el botón secundario del ratón en la tabla, en la Vista Diagrama, y aparece un menú que le permite crear una jerarquía (Figura 2-47).

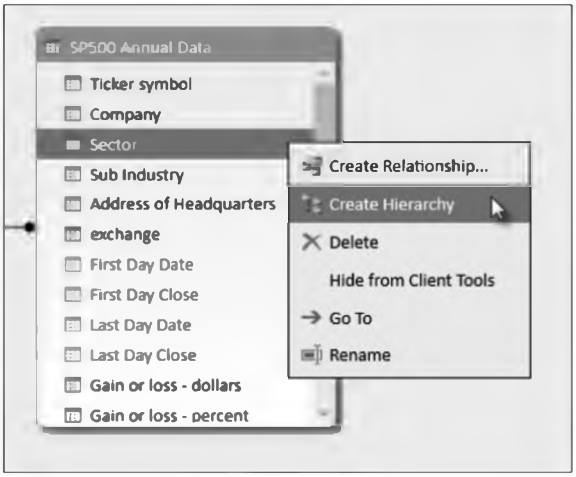


Figura 2-47

Se observa que la tabla SP 500 sigue teniendo las siglas GICS antes de Sector y Subsector. No se está creando la jerarquía en la tabla SP 500 (sino en la tabla Datos anuales de SP500), pero se quiere corregirlo de todas formas. Tras inspeccionarlo un poco más, se ve que los campos de la tabla Datos diarios de SP500 también tienen esas molestas siglas. Es posible corregirlo directamente desde la Vista Diagrama. Para ello se hace doble clic en el campo Sector GICS de la tabla SP 500 y se quitan las siglas GICS. Se hace lo mismo en los otros tres campos.

Ahora está todo listo para crear la jerarquía. Seleccione Datos anuales de SP500, nombre a la jerarquía Sector y subsectores e incluya el campo Sector y el campo Subsector de la tabla Datos anuales de SP500. El orden en el que están organizados los campos establece la jerarquía (Figura 2-48).

©Alfaomega-RC Libros

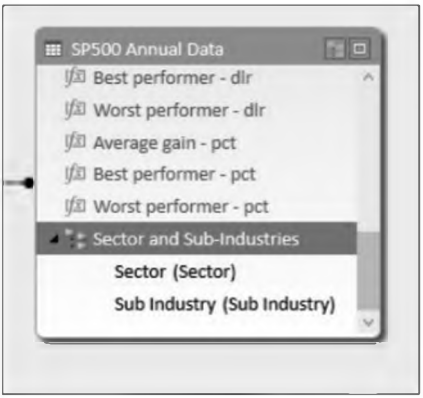


Figura 2-48

Con las relaciones establecidas y la jerarquía creada, está todo listo para comenzar a crear informes. Se quiere que los informes sean dinámicos e interesantes, con muchos elementos visuales. Y se quiere que sean interactivos para que, cuando se publiquen, sea posible revisar y analizar los datos de manera que resulten útiles (o interesantes). La función de BI con características de autoservicio que le permite hacer todo esto y más es Power View.

Hemos visto que en Power Pivot, se puede personalizar, ampliar con cálculos y jerarquías y administrar el eficaz modelo de datos que forma parte de Excel. Power Pivot funciona a la perfección y automáticamente con Power Query, y con otras características de Excel, por lo que le permite administrar y ampliar (con cálculos y jerarquías) su propia base de datos personalizada (el modelo de datos) en el entorno de Excel que ya conoce. Power Pivot incluye todos los datos que incorpore desde Power Query, más todos los datos que agregue al modelo. Además, todo el modelo de datos de Power Pivot (que incluye tablas, columnas, cálculos, jerarquías y todas las demás personalizaciones) se expone como elementos listos para crear informes en Power View.

CAPÍTULO 3

INFORMES CON POWER BI. VISUALIZACIÓN CON POWER VIEW Y POWER MAP

POWER VIEW

Con Power View, es posible crear visualizaciones interesantes, interactivas y exhaustivas con unos pocos clics. Se pueden obtener visualizaciones de todo tipo y los informes creados en Power View se comparten fácilmente con otros usuarios. Los consumidores de informes pueden analizarlos de forma interactiva.

Para crear un nuevo informe de Power View en Excel, se elige el botón *Power View*, que se encuentra en la sección *Informes* de la cinta de opciones *Insertar* de Excel (Figura 3-1). Se crea una nueva pestaña con una hoja de cálculo.

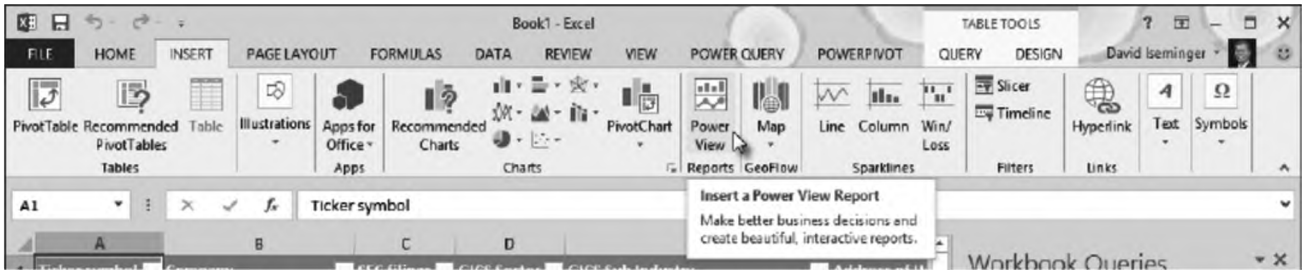


Figura 3-1

La pestaña *Power View* solo aparece en la cinta de opciones cuando el informe de Power View es la hoja de cálculo activa. Con la nueva hoja de informe de Power View generada, aparece la pestaña Power View en la cinta de opciones y ya se tiene todo lo necesario para crear un informe.

Si se activó la casilla *Cargar en libro* en Power Query, verá tablas adicionales en el panel *Campos* de Power View, una por cada consulta que se cargó en el libro. Power View incluye todas las tablas del modelo de datos y el libro en *Campos* de Power View. Puede volver a Power Query y desactivar la casilla *Cargar en libro* para quitar esa tabla del libro y de *Campos* de Power View (Figura 3-2).

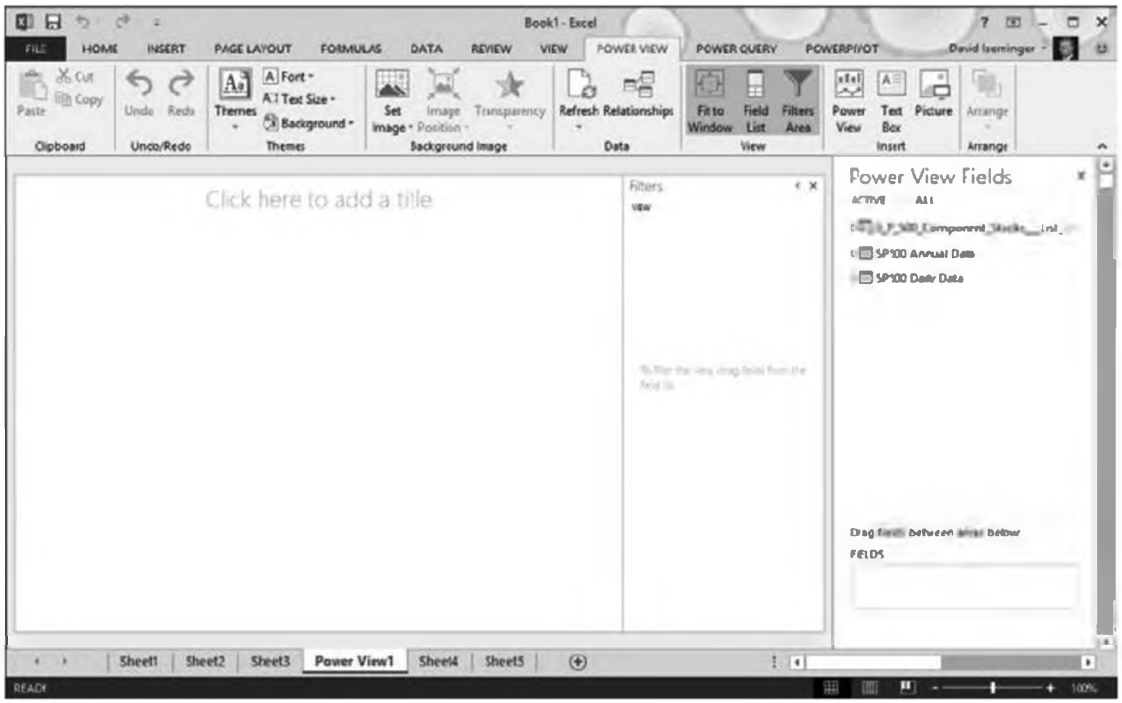


Figura 3-2

En nuestro primer informe queremos mostrar los resultados de cada sector por separado, y una comparativa de los sectores en cada bolsa.

©Alfaomega-RC Libros

Se empieza por ir a la tabla Datos anuales de SP500 del panel *Campos* de Power View y, a continuación, se hace clic en la flecha hacia abajo para ver los campos disponibles (Figura 3-3).

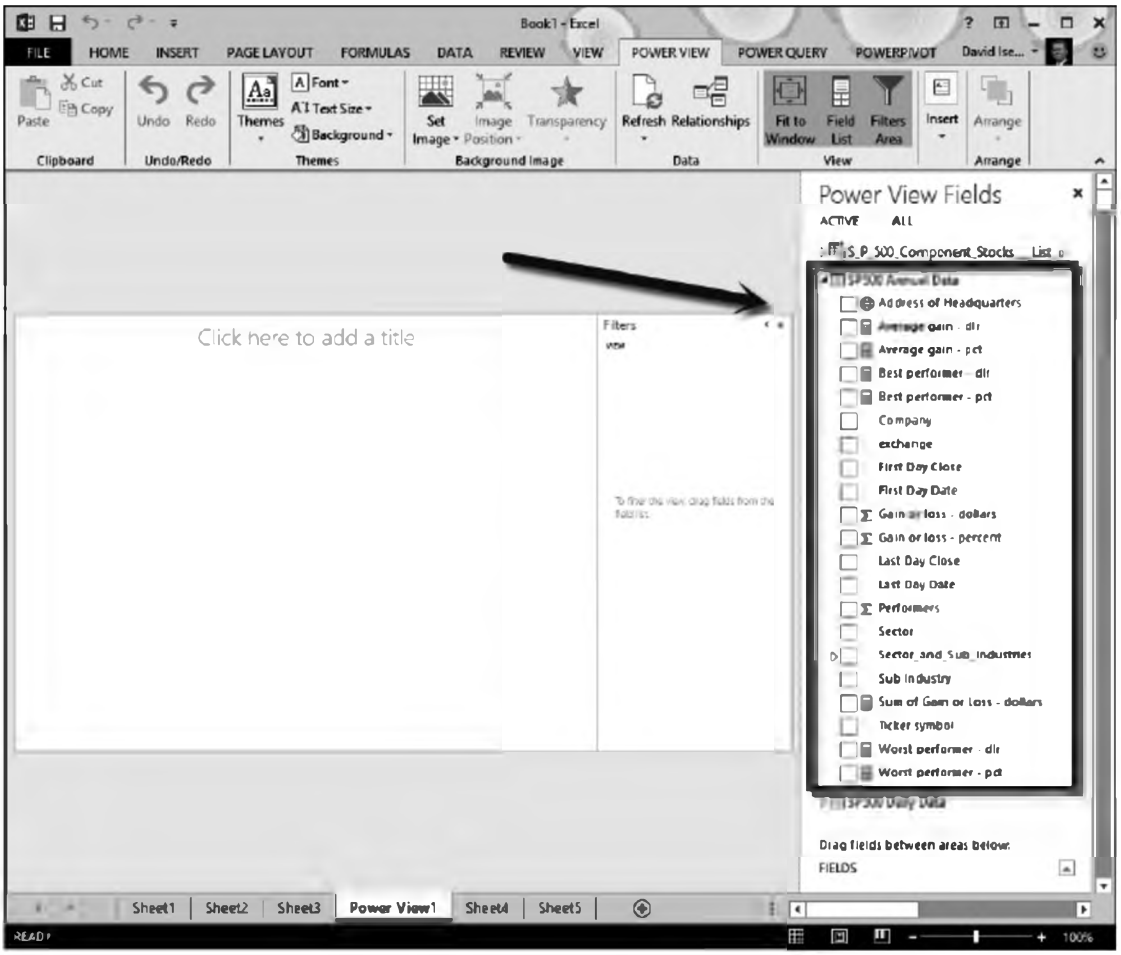


Figura 3-3

Hay muchísimos campos, muchos de los cuales no se necesitan para crear los informes. Puede que se pasara un poco con los campos calculados que creó en Power Pivot. Resulta un poco confuso tener todos esos campos disponibles. No se quiere buscar entre tantos campos cada vez que se vaya a agregar uno a una visualización. Para solucionar este problema, se pueden ocultar algunas columnas, o campos, y así evitar que aparezcan en la tabla al crear los informes en Power View. Elija el botón *Administrar* de la cinta de opciones Power Pivot para iniciar Power Pivot. Seleccione la tabla que contiene los Datos anuales de SP500, haga clic con el botón secundario del ratón en la columna quiere ocultar y elija *Ocultar en herramientas cliente* (Figura 3-4).

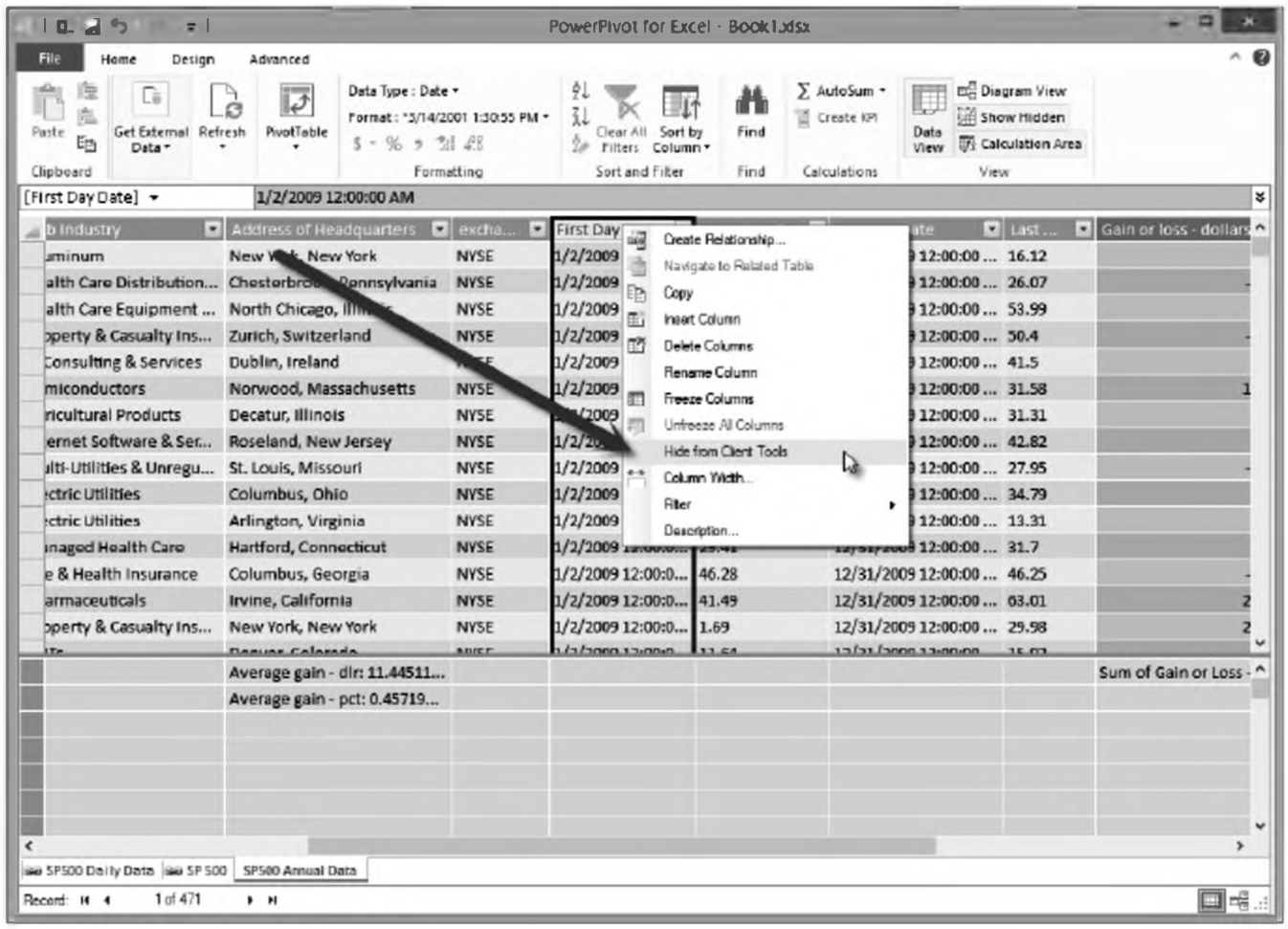


Figura 3-4

Los datos subyacentes siguen en la tabla (y en el modelo de datos), pero el campo se ocultará en las herramientas del cliente, como Power View. Se ocultan varias columnas más que no se necesitan para crear informes. Cuando una columna está oculta en las herramientas del cliente, aparece en gris en Power Pivot. También es posible ocultar cálculos en el área *Campos calculados*. Se pueden mostrar u ocultar los campos ocultos en la ventana *Administrar* de Power Pivot seleccionando *Inicio -> Ver -> Mostrar ocultos* en la cinta de opciones (Figura 3-5).

Se ocultan los siguientes campos en la tabla Datos anuales de SP500:

- Fecha del primer día
- Cierre del primer día
- Fecha del último día
- Cierre del último día
- Valores con mejores y peores resultados
- Los cálculos de los valores con los mejores y peores resultados (cuatro de ellos)

©Alfaomega-RC Libros

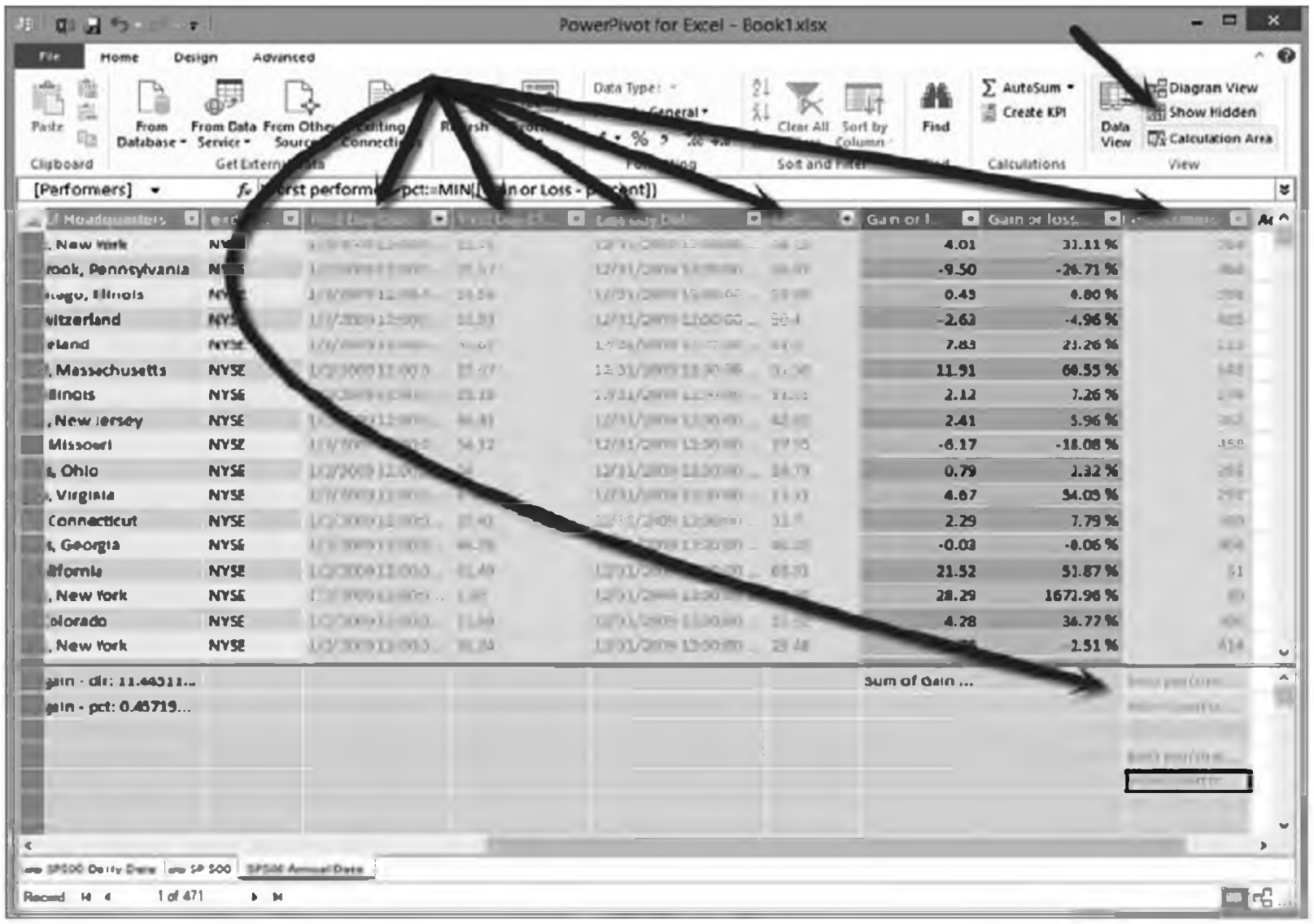


Figura 3-5

Cuando regresamos a Excel, se ve una notificación que indica que el modelo de datos ha cambiado (Figura 3-6). Esta notificación aparece cada vez que se hacen modificaciones en el modelo de datos, ya sea ocultar campos, crear nuevos cálculos o quitar columnas personalizadas.

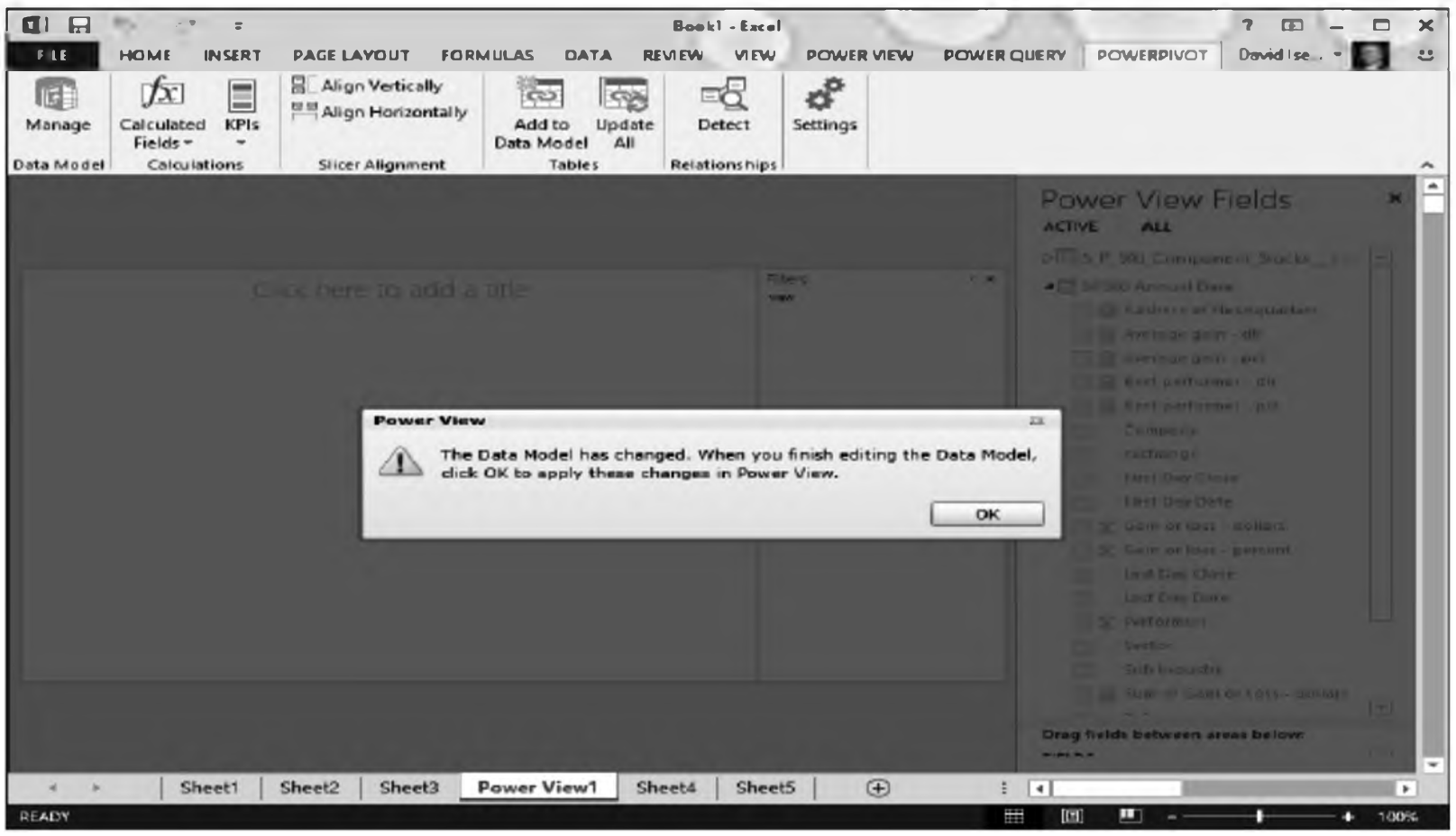


Figura 3-6

Esta posibilidad significa que se podrá continuar modificando y personalizando el modelo de datos, incluso después de pensar que se ha creado el conjunto de datos correcto y la mejor colección posible de campos calculados. Puede personalizar todo lo que quiera al crear los informes, y Excel se asegurará de que el modelo de datos y todos los informes estén sincronizados y actualizados.

Elija *Aceptar* y revise la lista de campos disponibles de la tabla Datos anuales de SP500 (Figura 3-7). Se decide que los campos disponibles son mucho más fáciles de manejar si son menos.

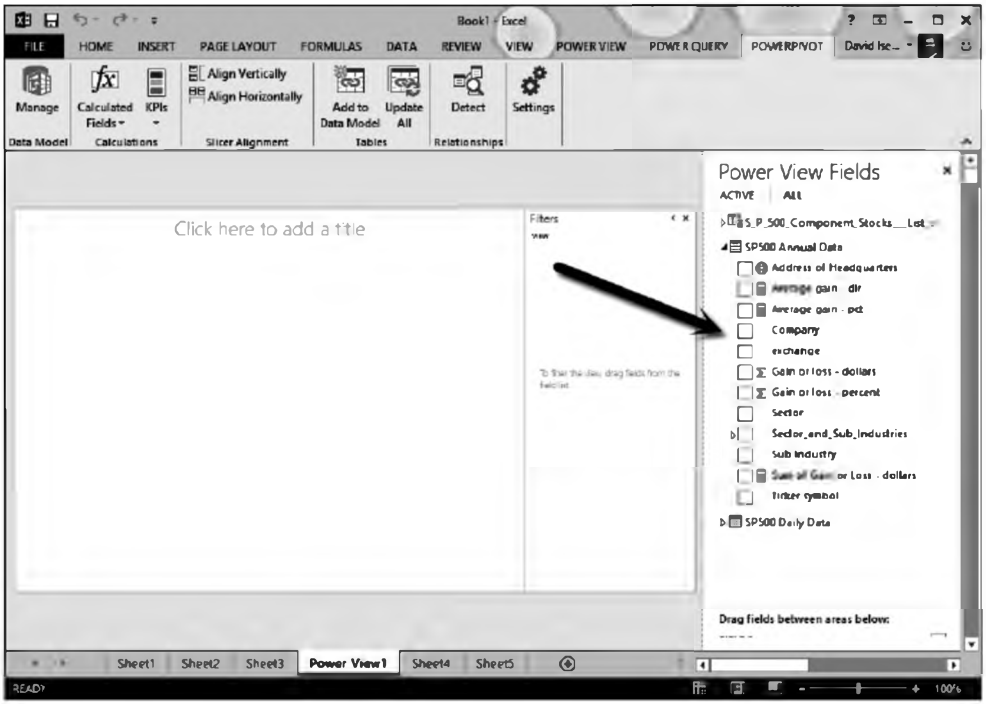


Figura 3-7

Si necesita un campo que está oculto, puede volver a Power Pivot y mostrar el campo con la misma facilidad, siguiendo los mismos pasos. Cuando una columna o un cálculo que está oculto está seleccionado en Power Pivot, al hacer clic con el botón secundario, la opción del menú pasa a ser Mostrar en herramientas cliente.

En cada uno de los siguientes informes, se aprende algo sobre Power View que nos permite crear informes más fácilmente, con más eficacia y de forma que sean más interesantes para los consumidores de esos informes.

©Alfaomega-RC Libros

Informes sencillos

Para empezar con el primer informe, se arrastra el campo Ganancia o pérdida en dólares del panel *Campos* de Power View hasta el lienzo del informe. Se crea una tabla. Luego, se arrastra el campo *Sector* a la tabla que acaba de crear. Power View crea la tabla siguiente (Figura 3-8).

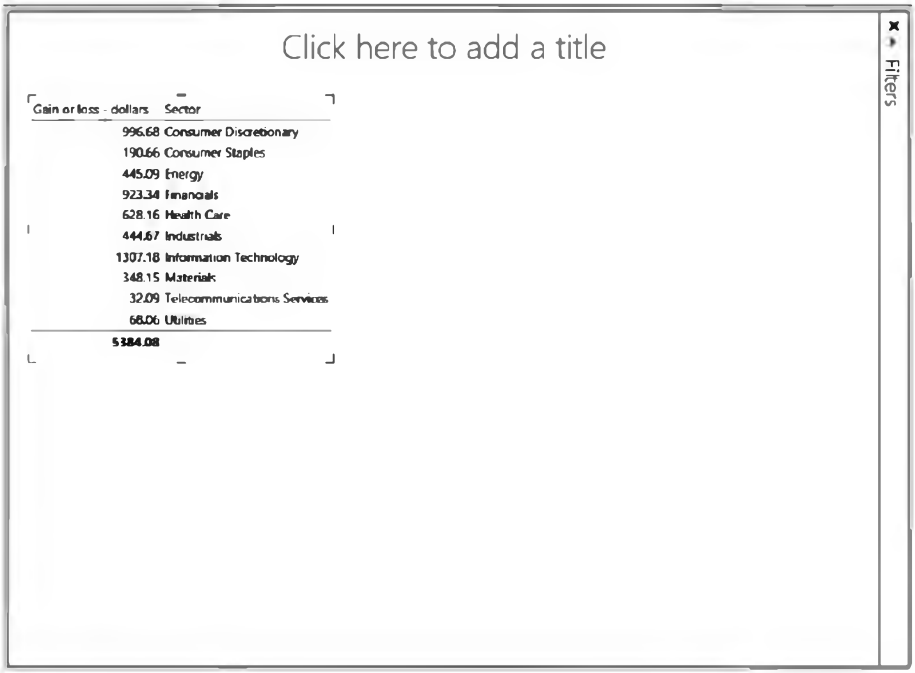


Figura 3-8

Power View crea automáticamente una suma de las ganancias. Se decide que sería más interesante ver la ganancia promedio de cada sector, al menos en este informe, así que se agrega ese campo. Para ver el resultado como un gráfico de barras, se selecciona la tabla y, después, se elige *Diseño -> Cambiar visualización -> Gráfico de barras -> Barra apilada* en la cinta de opciones (Figura 3-9).

Ya se está consiguiendo algo. Pero todas son del mismo color, y se quiere que cada sector tenga su propio color (Figura 3-10).

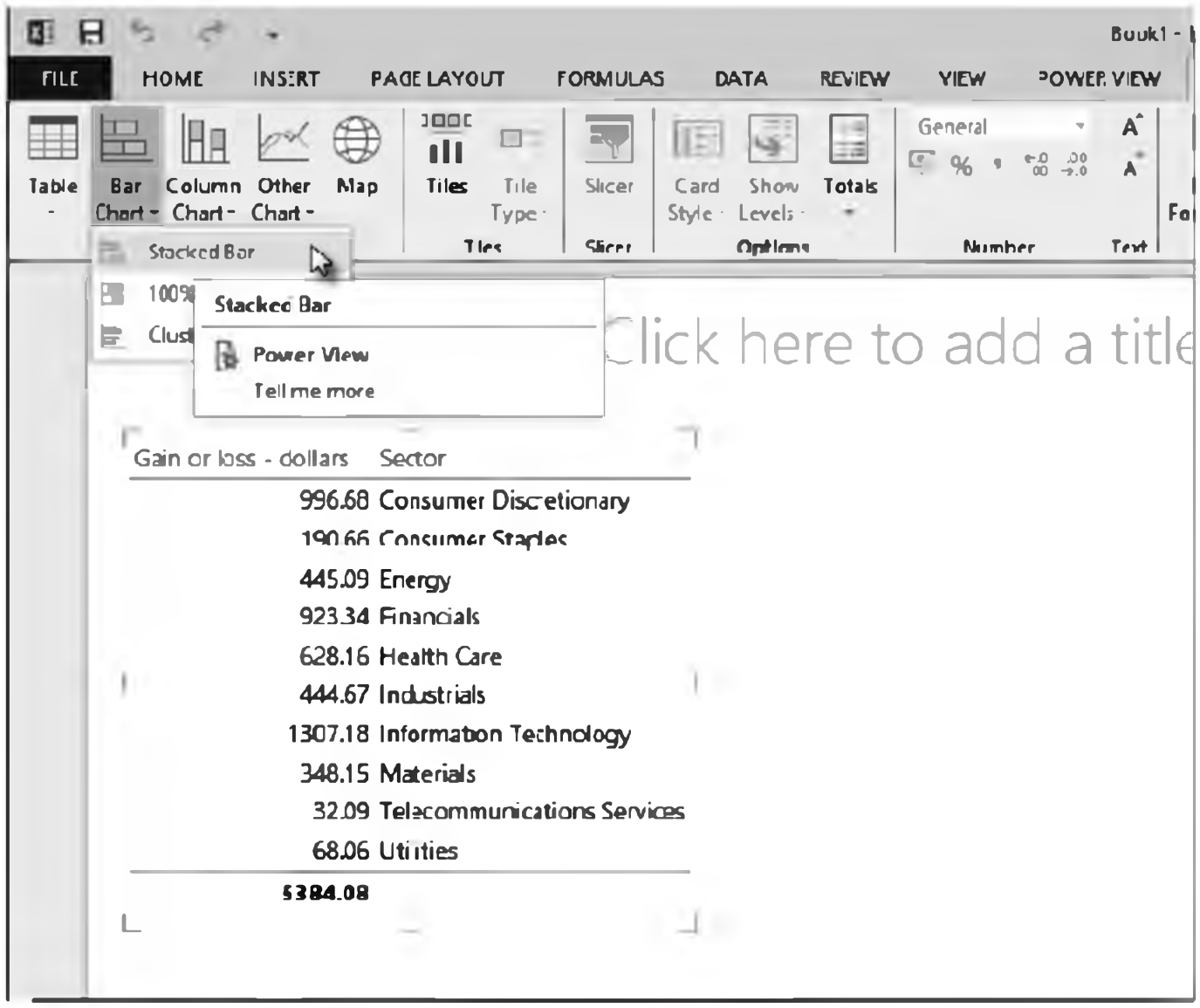


Figura 3-9

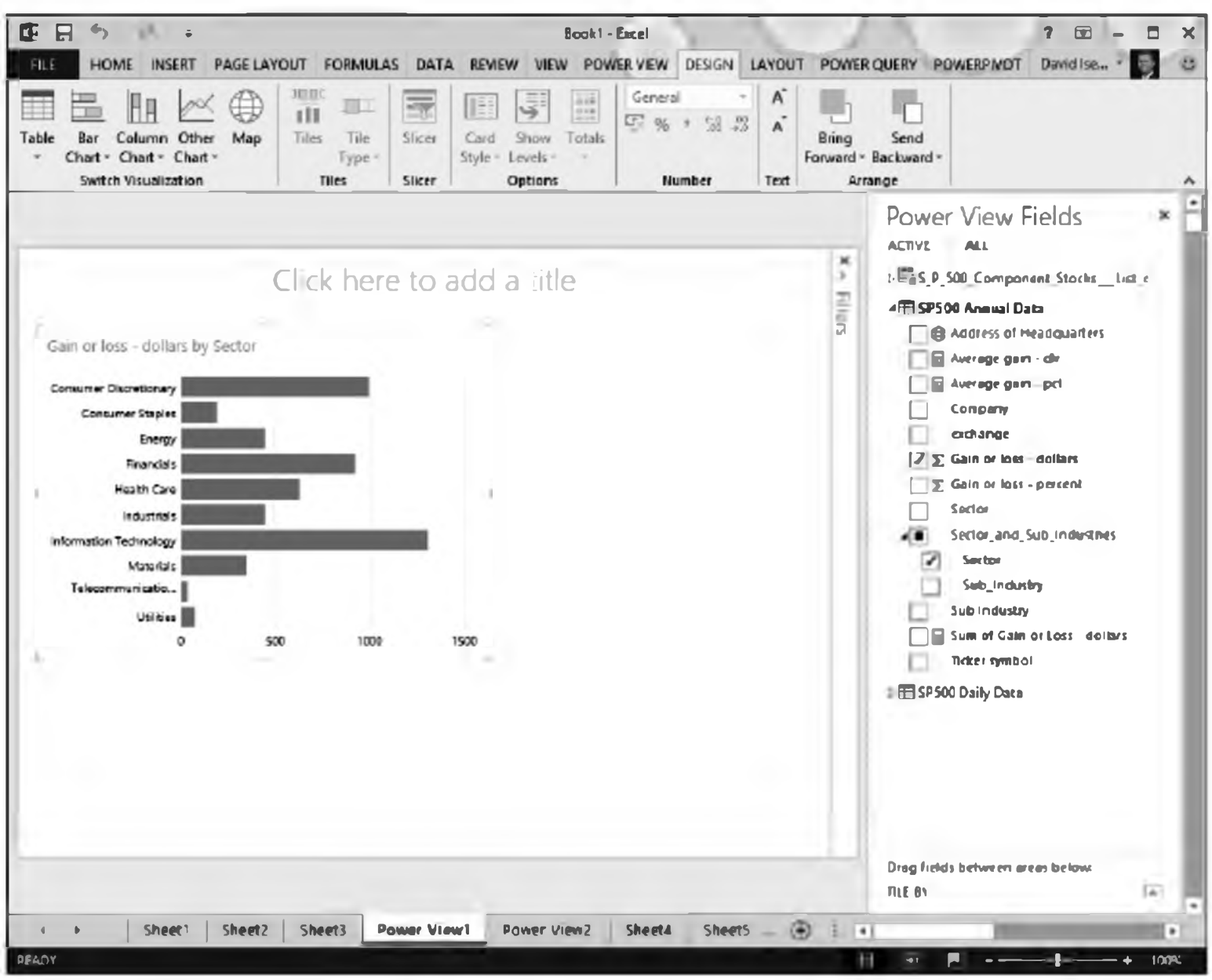


Figura 3-10

©Alfaomega-RC Libros

Queremos cambiar la leyenda del gráfico de barras apiladas. Se desea que cada sector esté de un color diferente, por lo que se selecciona la visualización de gráfico de barras y, esta vez, se arrastra el campo *Sector* hasta la sección *Leyenda* de *Campos* de Power View (Figura 3-11).

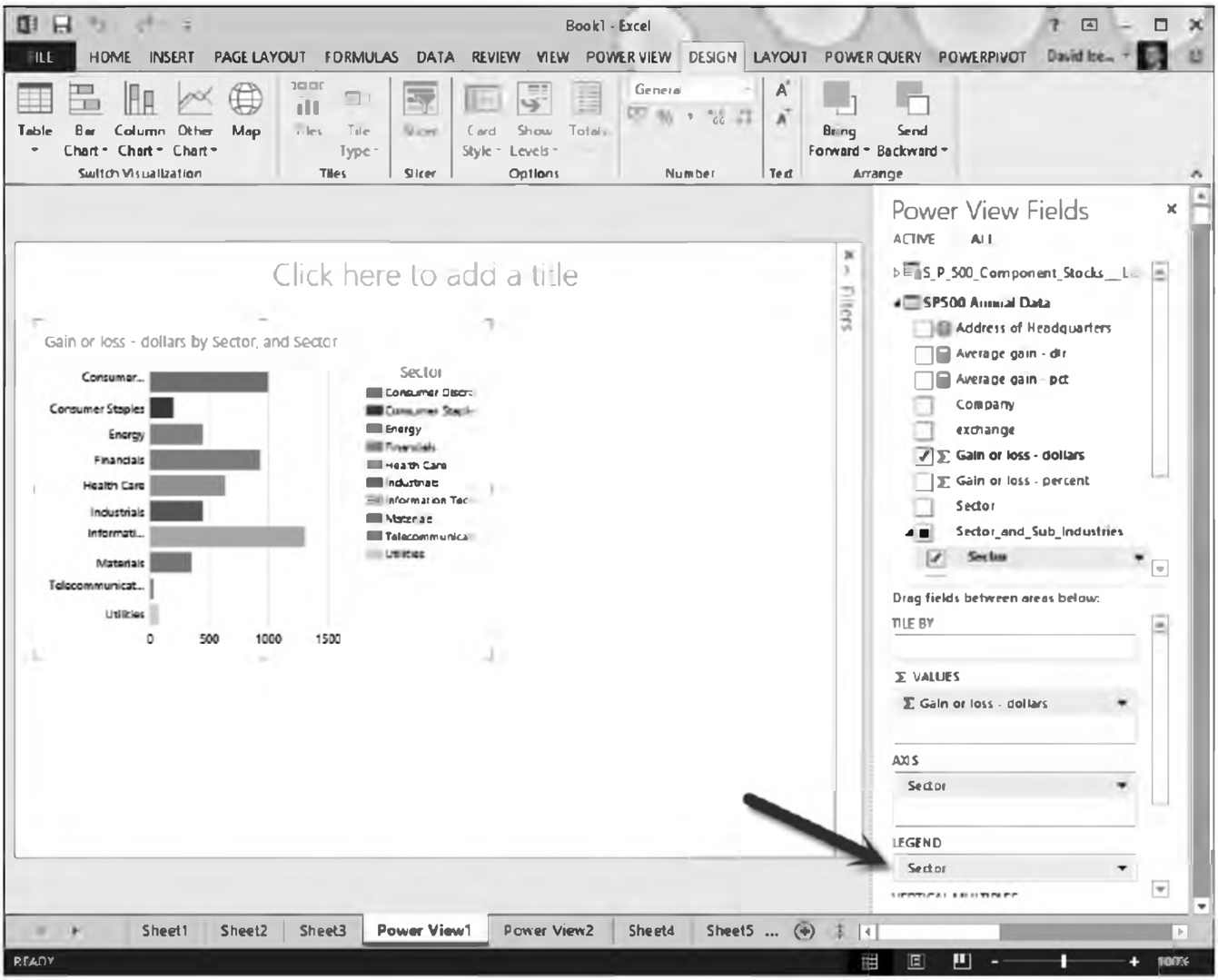


Figura 3-11

Ya se tiene el nombre del sector a la izquierda de cada barra, por lo que no se necesita mostrar la leyenda en la visualización. Para quitar esa leyenda, con el gráfico seleccionado, se elige *Diseño -> Etiquetas -> Leyenda* y, a continuación, se elige *Ninguna* (Figura 3-12).

Los colores del gráfico de barras siguen usando Sector como leyenda, pero la leyenda en sí no se muestra en la visualización (Figura 3-13).

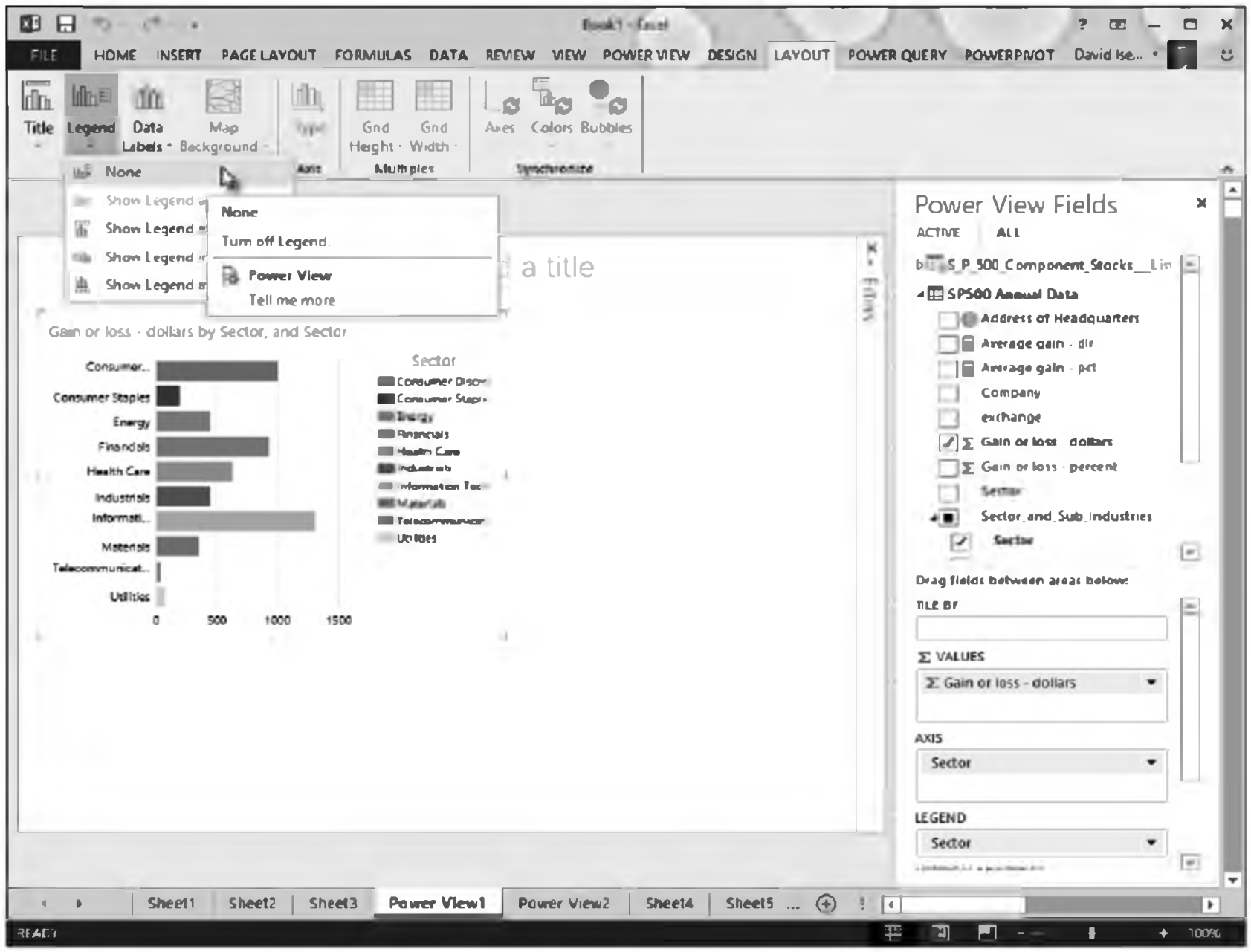


Figura 3-12

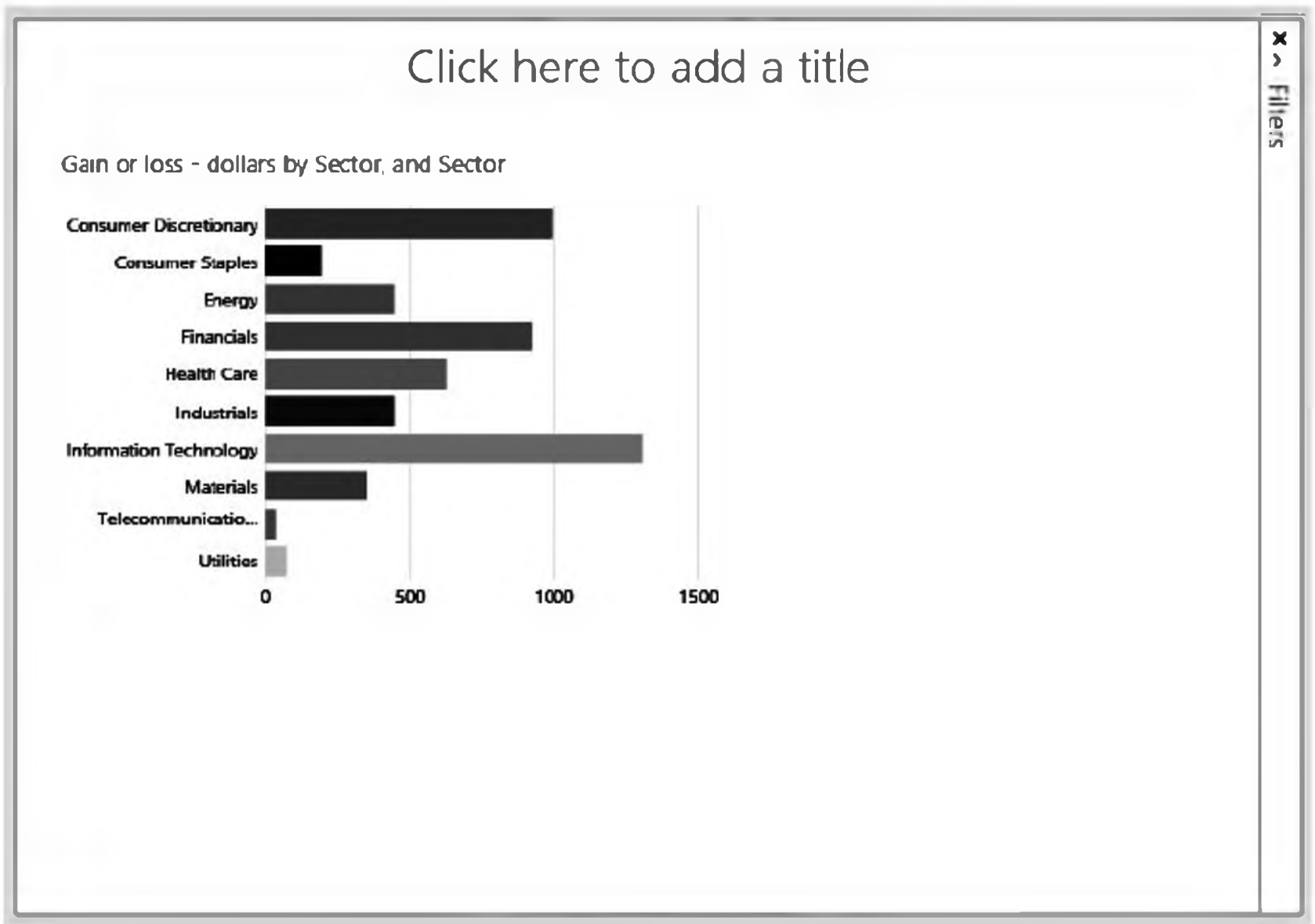


Figura 3-13

©Alfaomega-RC Libros

Ya sabemos crear gráficos de este informe. Se crean algunos más en esta página del informe para resaltar cada sector, además de sus resultados anuales y a lo largo del año. También se escribe el título Resultados de los sectores en el propio informe. Asimismo, para mantener ordenadas las pestañas del libro, se cambia el nombre de la pestaña a Resultados de los sectores. Cuando la página se muestra por primera vez, se muestran todos los sectores en todas las visualizaciones (Figura 3-14).

Cuando se hace clic o se pulsa en un sector del gráfico de barras, todo el informe se filtra para mostrar únicamente ese sector, y todas las visualizaciones muestran los resultados de ese sector. En la Figura 3-15 tenemos el sector de productos de consumo básico.

Observe que el color de cada gráfico de líneas también refleja el color del sector seleccionado en el gráfico de barras. Esto se debe a que se seleccionó el mismo campo (Sector) para la leyenda de todos los gráficos de líneas.

En la Figura 3-16 se observa cómo aparece el sector de la informática cuando se selecciona en el gráfico de barras.

Figura 3-14

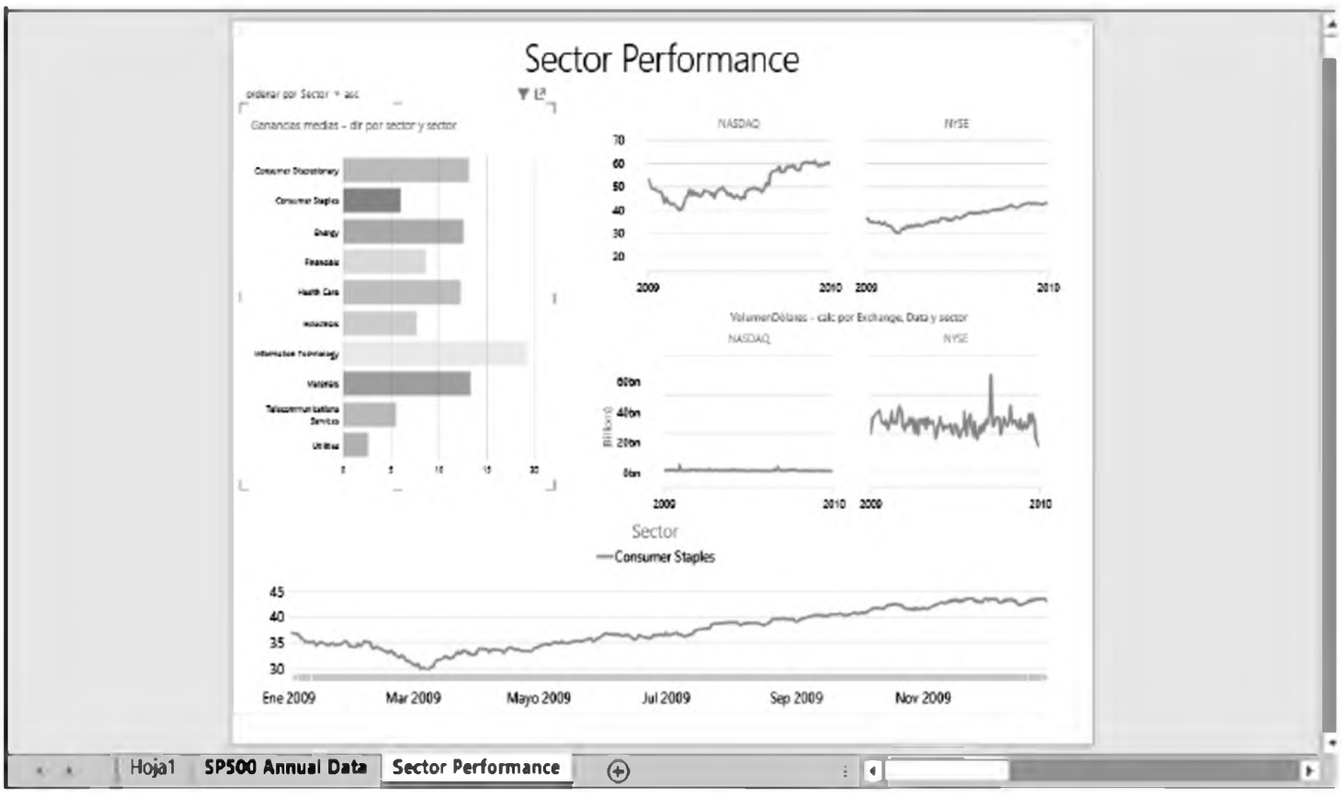


Figura 3-15

Figura 3-16

©Alfaomega-RC Libros

En la Figura 3-17 se presenta el sector de la energía. Observe que el gráfico de la esquina superior derecha, que muestra el volumen en dólares diario de los títulos negociables del sector, ya no tiene ningún gráfico de NASDAQ. ¿Por qué? Después de investigar un poco, se ve que es porque no hay ningún título negociable del NASDAQ en el sector energético, de modo que Power View quita automáticamente ese gráfico de la visualización. No hay títulos de energía en NASDAQ.

Figura 3-17

Sabemos que todas las personas que vean este informe también podrán seleccionar un sector, y el informe mostrará los resultados filtrados, igual que lo que vemos nosotros. Será una herramienta de presentación muy útil, piensa.

Informes detallados

Ahora se quieren ver más detalles de uno de los sectores. Este informe tiene un formato similar, pero incluye una capa más profunda en la jerarquía de Sector: Subsector. Se arrastra el campo *Subsector* de la jerarquía de Sector hasta el lienzo del informe. Se debe señalar que al crear una jerarquía, todos los niveles de la jerarquía siguen estando disponibles individualmente para crear informes (Figura 3-18).

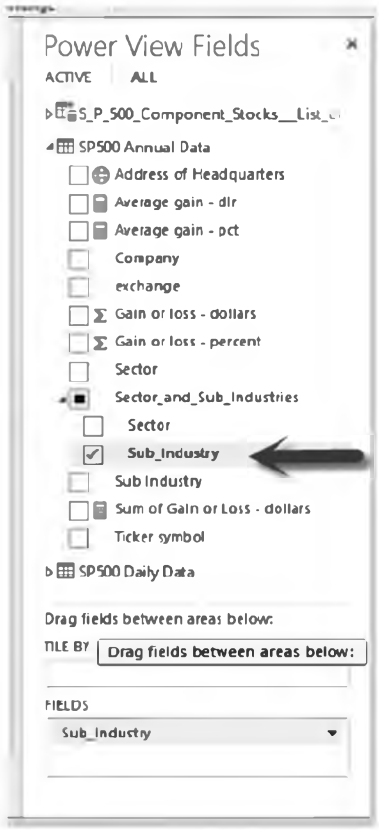


Figura 3-18

Por ejemplo, en este segundo informe, se arrastró el campo Subsector hasta el lienzo. Eso funcionó en un principio, pero hay demasiados subsectores como para mostrarlos todos. Y, en realidad, solo nos interesa el sector de productos de consumo básico.

Observamos, en la Figura 3-19, que Power View agrega una barra de desplazamiento a la tabla. Así, se puede ver el resto de los datos de la tabla. Observe también el panel *Filtros* de Power View. Este se puede usar para filtrar la Tabla seleccionada o toda la Vista (todo el informe de Power View).

Ahora se quiere que todo el informe contenga solo los datos del sector de productos de consumo básico, por lo que se arrastra *Sector* a la sección *Vista* del panel *Filtros*. Luego, se hace clic en la flecha hacia abajo que se encuentra junto a *Sector* y se selecciona *Productos de consumo básico* (Figura 3-20).

©Alfaomega-RC Libros

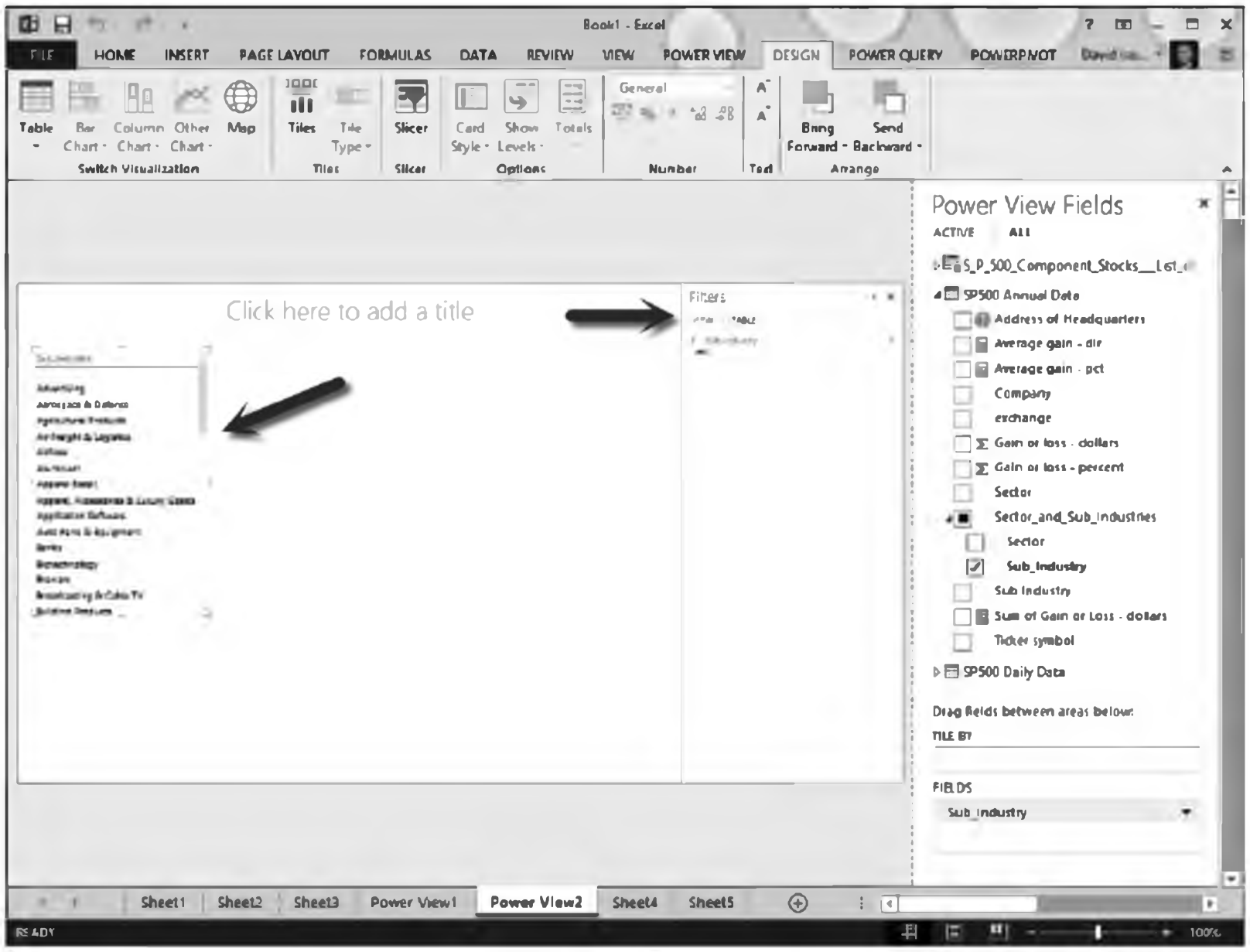


Figura 3-19

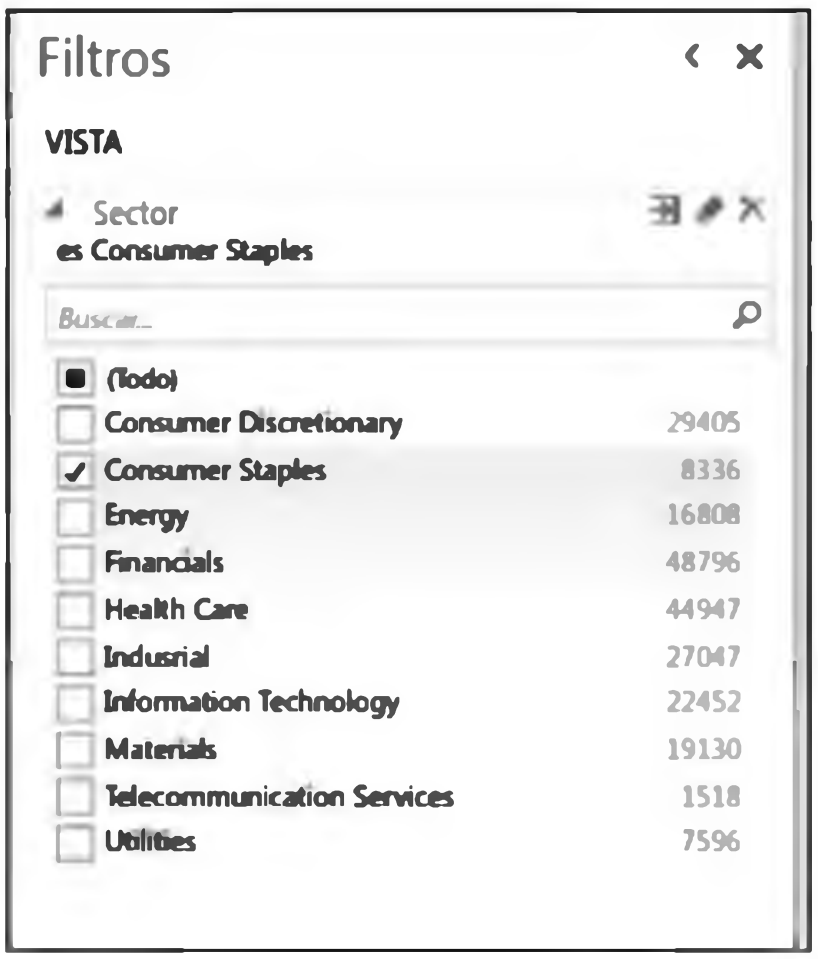


Figura 3-20

Vamos a crear unas cuantas visualizaciones más para obtener un informe que analice los resultados de los subsectores del segmento Productos de consumo básicos. Cambie el nombre de la pestaña Power View de Excel a Productos de consumo básico. Como en el primer informe, al seleccionar uno de los subsectores en el gráfico de barras, se filtran las visualizaciones en el resto del informe (Figura 3-21).

Figura 3-21

Ahora se activa otro subsector para comprobar que los gráficos se comportan de la manera que quiere. Se muestran justo como se esperaba (Figura 3-22).

©Alfaomega-RC Libros

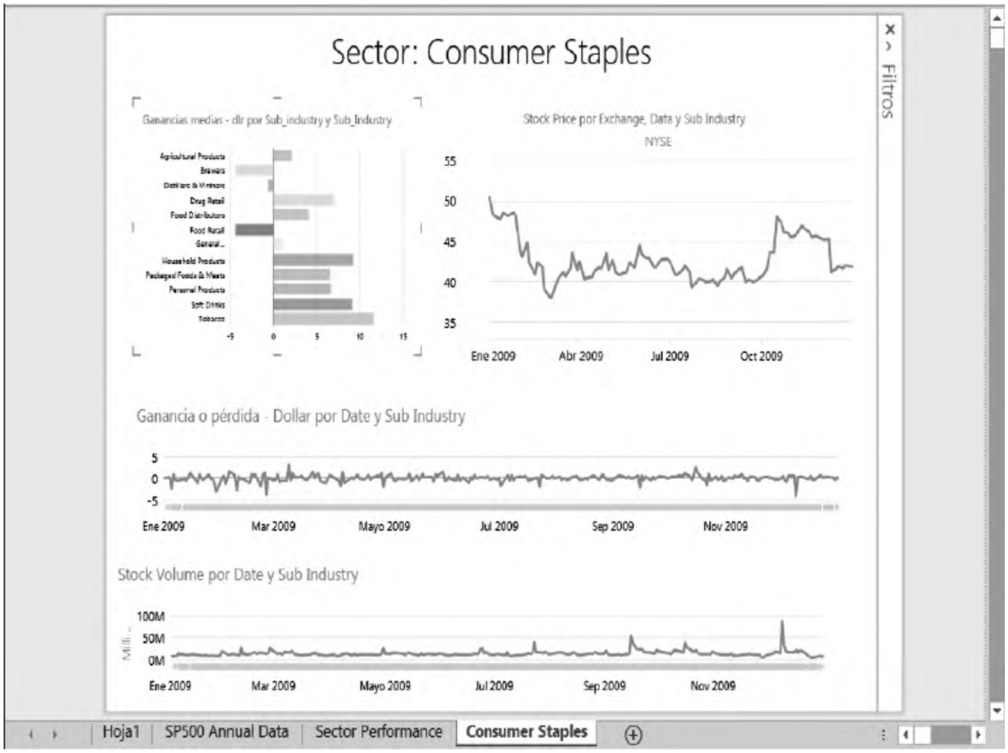


Figura 3-22

Informes con segmentaciones de datos

Estamos bastante satisfechos con el informe de productos de consumo básicos, pero no es apetecible crear un informe de Power View diferente para cada sector. Probablemente hay una manera de obtener todos los sectores, y todos los subsectores asociados a ellos, en un solo informe.

Se puede hacer con segmentaciones de datos. Las segmentaciones de datos de Power View se comportan de forma similar a las segmentaciones de datos de Excel: filtran todas las visualizaciones a la vez. Para crear una segmentación de datos, se arrastra el campo con el que quiere segmentar los datos (en este caso, el campo *Sector* de la tabla Datos anuales de SP500) hasta el lienzo. Se crea una tabla de Power View basada en ese campo. Después, con la tabla seleccionada, se elige *Diseño -> Segmentación de datos -> Segmentación de datos* en la cinta de opciones (Figura 3-23).

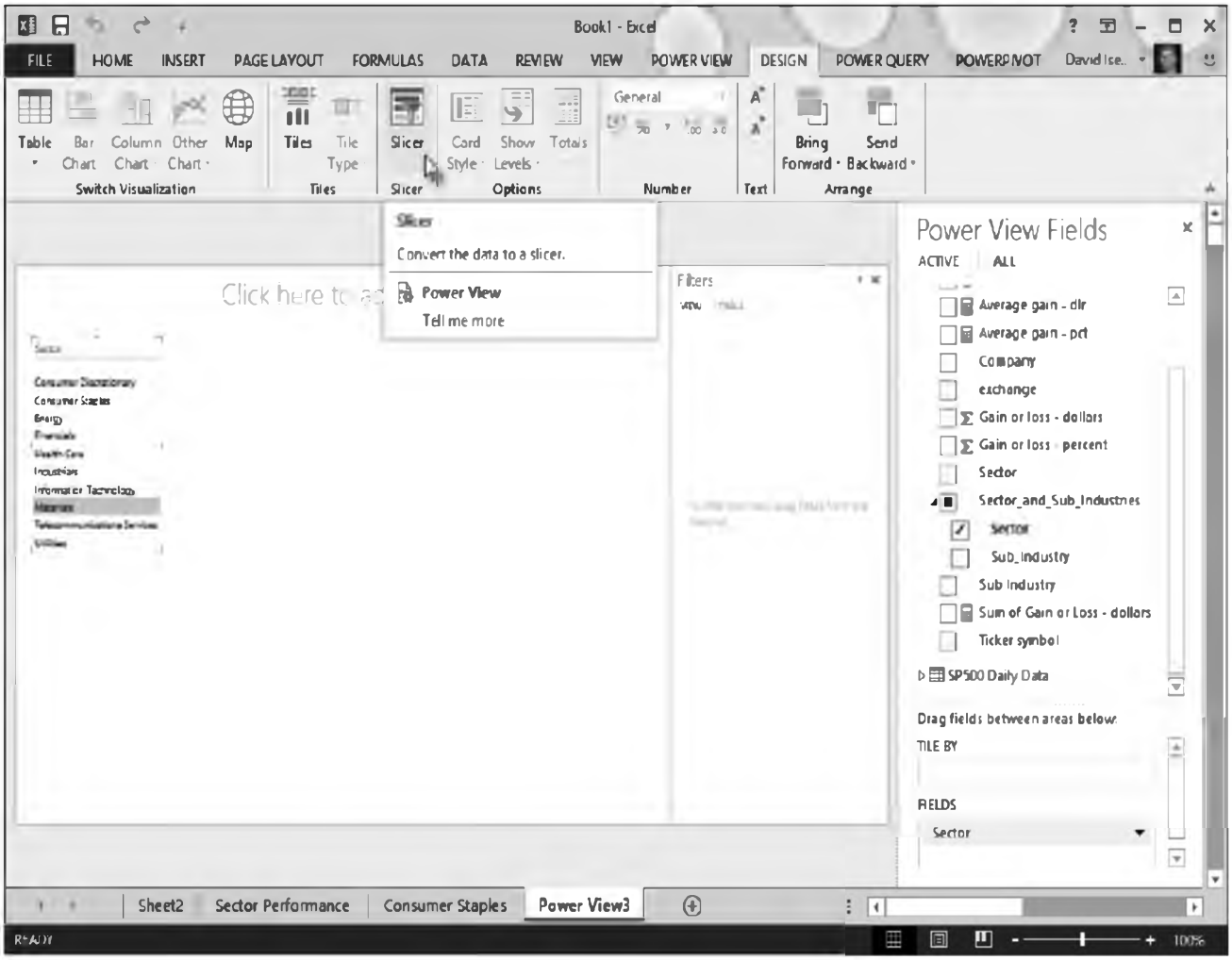


Figura 3-23

Se crean unas cuantas visualizaciones más, incluido un gráfico de barras que mostrará cada subsector y permitirá a los consumidores de informes seleccionar un sector y luego un subsector para visualizar sus resultados. Cuando se termina el informe, se nombra como Resultados de subsectores (Figura 3-24).

Cuando los que ven el informe interactúen con el informe, podrán seleccionar un sector en la segmentación de datos (parte superior izquierda) y, luego, seleccionar un subsector en el gráfico de barras (parte inferior izquierda), y los gráficos de líneas diarios reflejarán los resultados de ese subsector.

Por ejemplo, la siguiente pantalla muestra el sector Sanitario y su subsector Distribuidores y servicios sanitarios (Figura 3-25).

©Alfaomega-RC Libros

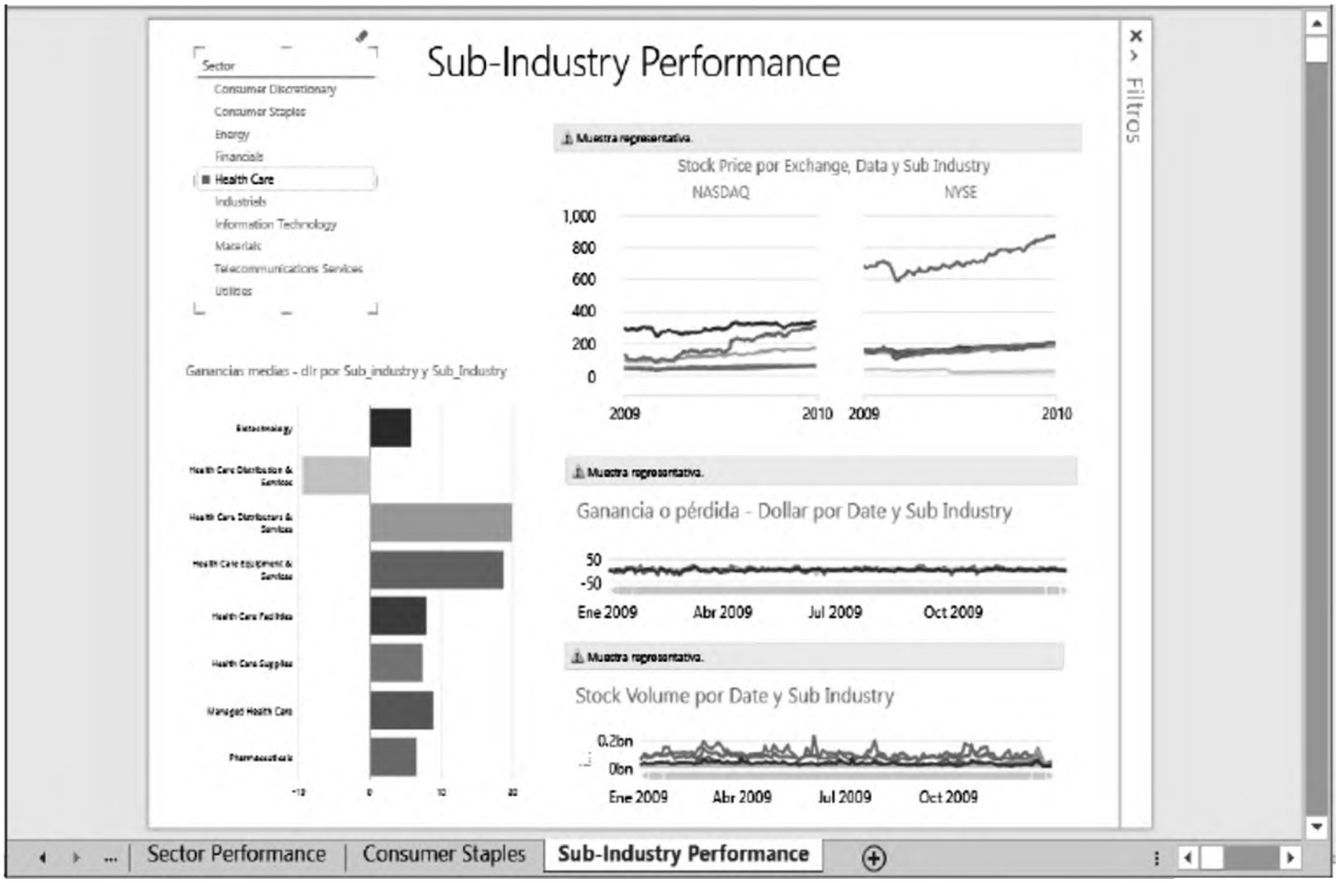


Figura 3-24

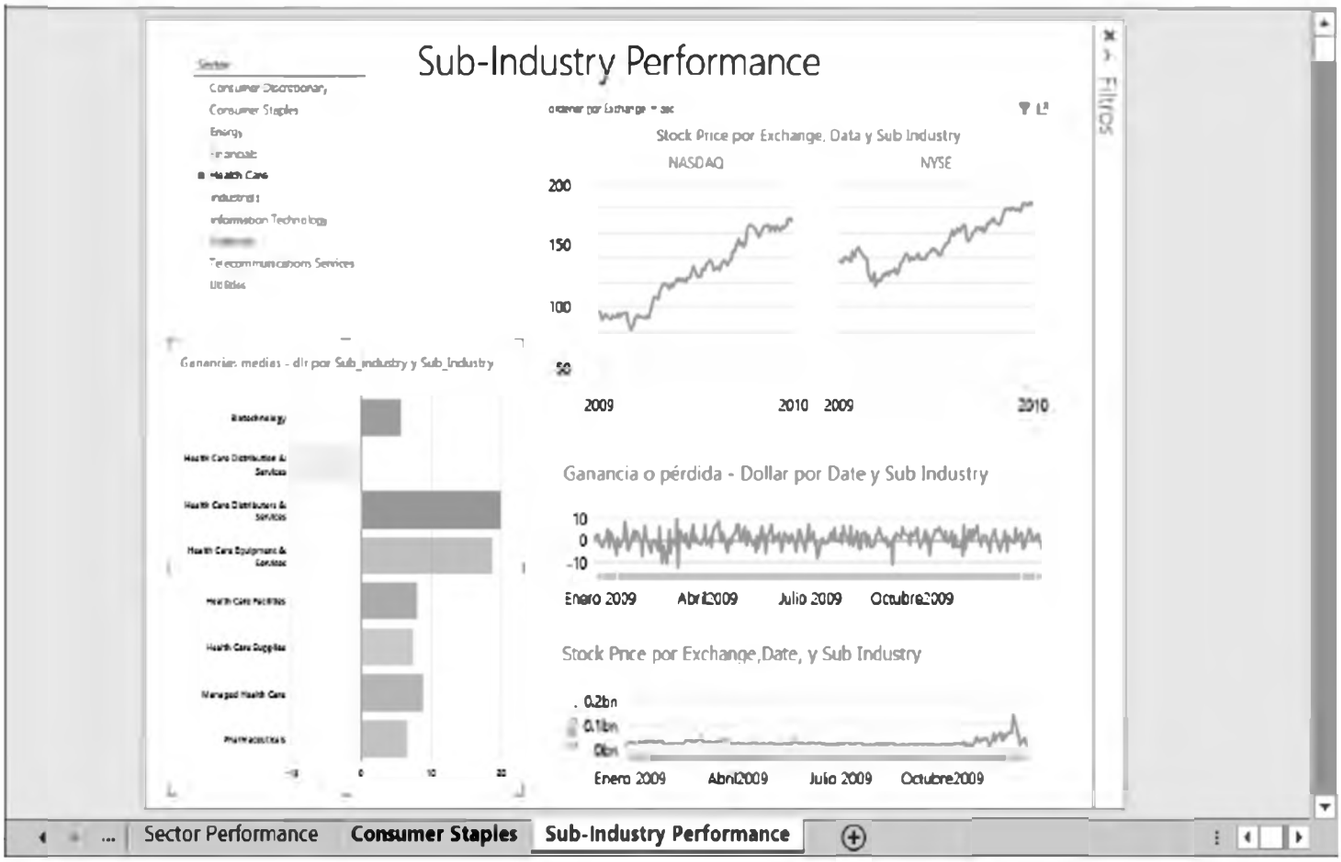


Figura 3-25

Aunque este informe es de calidad, sería interesante que cualquier analista pudiese explorar en profundidad un sector determinado, observar los números y compararlos con otros. También sería interesante comparar el crecimiento promedio en dólares con la ganancia promedio en porcentaje uno al lado del otro. Intentaremos ampliar el informe de Power View para hacer comparaciones.

Informes de comparaciones

Vamos a considerar un informe que utilice la jerarquía que ya se creó previamente en Power Pivot. Creamos un nuevo informe de Power View, esta vez eligiendo *Power Pivot -> Insertar -> Informe*. Arrastramos la jerarquía de sectores de la tabla Datos anuales de SP500 hasta el lienzo.

Agregamos *Sector* al área *Leyenda de Campos* de Power View, luego *Ganancia promedio en dólares* y, a continuación, cambiamos la visualización a un gráfico de barras apiladas. La imagen de la Figura 3-26 muestra el resultado.

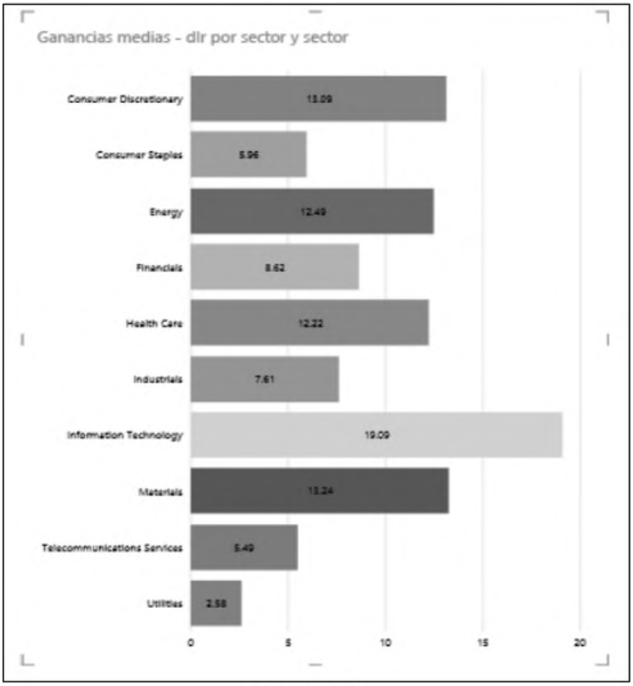


Figura 3-26

©Alfaomega-RC Libros

Creamos una tabla similar para Ganancia promedio en porcentaje y las alineamos una al lado de la otra. Llamamos a la pestaña de Power View y al informe Comparaciones de sectores y titulamos el informe Comparaciones de sectores con desgloses. Creemos que las visualizaciones serían más útiles si cada barra incluyera también un valor. Utilizamos *Diseño -> Etiquetas -> Etiquetas de datos* y elige *Mostrar* (Figura 3-27). Si agregamos *Etiquetas de datos* a ambos informes, el informe de Power View tendrá el aspecto de la Figura 3-28.

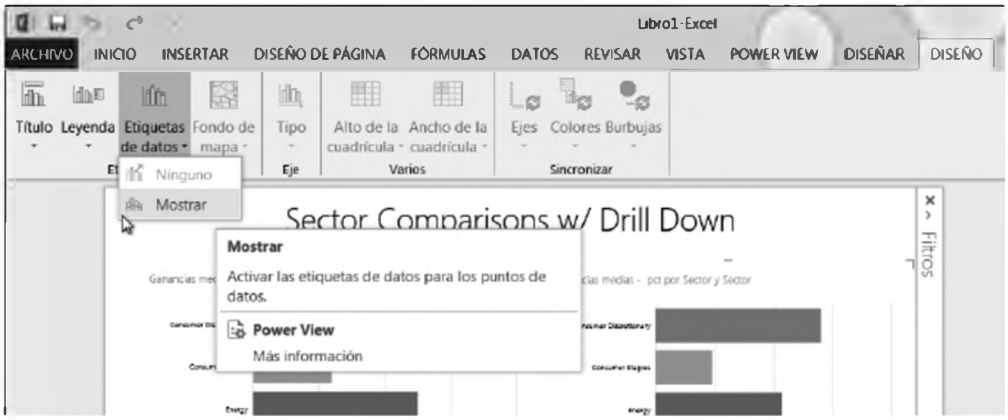


Figura 3-27

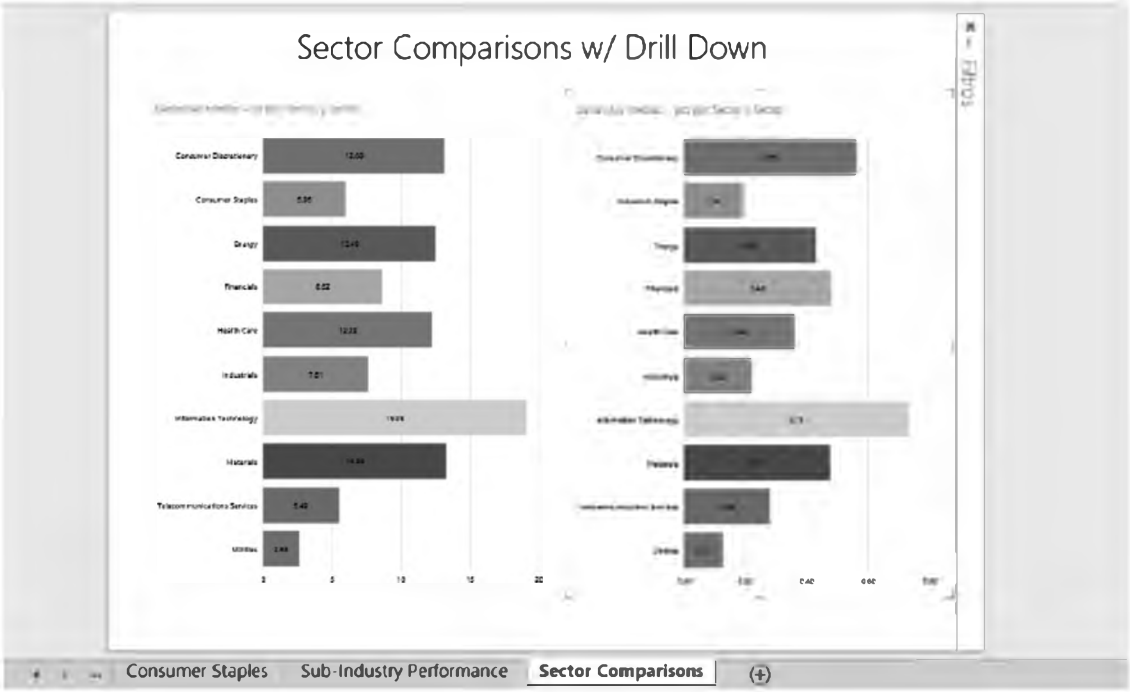


Figura 3-28

Al igual que otros informes, los gráficos permanecen sincronizados al seleccionar un sector. Y al mantener el ratón sobre un punto de datos, aparece un banner que muestra información sobre ese punto de datos, como se muestra en la Figura 3-29.

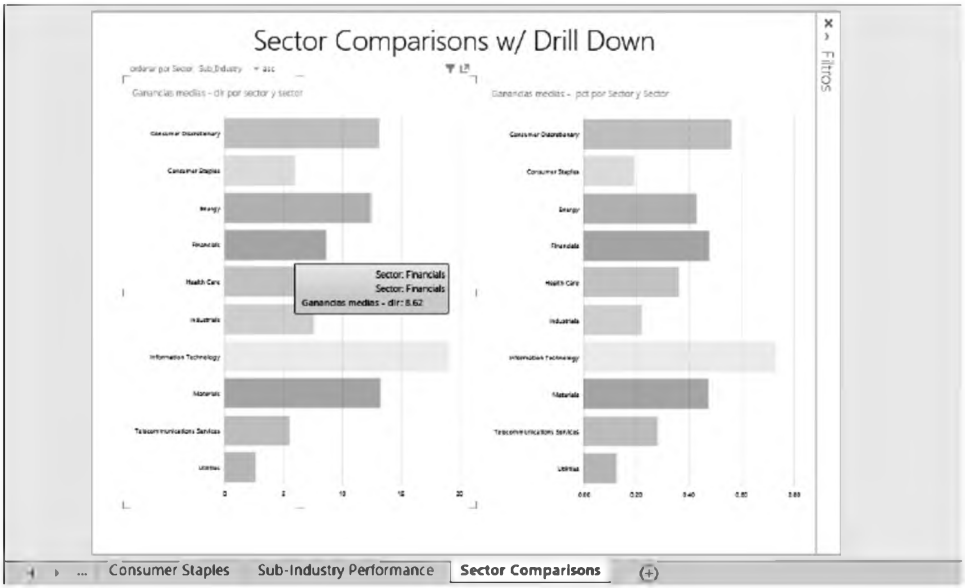


Figura 3-29

Pero lo más interesante de este gráfico es la capacidad de explorar en profundidad hasta ver los datos jerárquicos. Puesto que creamos una jerarquía y la incluimos en la visualización, todas las personas que vean este informe pueden hacer doble clic en cualquier sector y explorar en profundidad los detalles del subsector (Figura 3-30).

Al explorar en profundidad el sector financiero y observar las sociedades de inversión inmobiliaria minorista, se descubre que en medio de la crisis mundial, que comenzó con las hipotecas de alto riesgo y una caída de los valores del sector inmobiliario, este sector (al menos los tres subsectores del sector financiero, a continuación) creció más que la mayoría (Figura 3-31).

Cuando se finaliza y se quiere volver a un nivel más general, o regresar un nivel en la jerarquía, se elige el icono obtención de detalles ascendente, que se encuentra en la esquina superior derecha de la visualización (Figura 3-32).

©Alfaomega-RC Libros

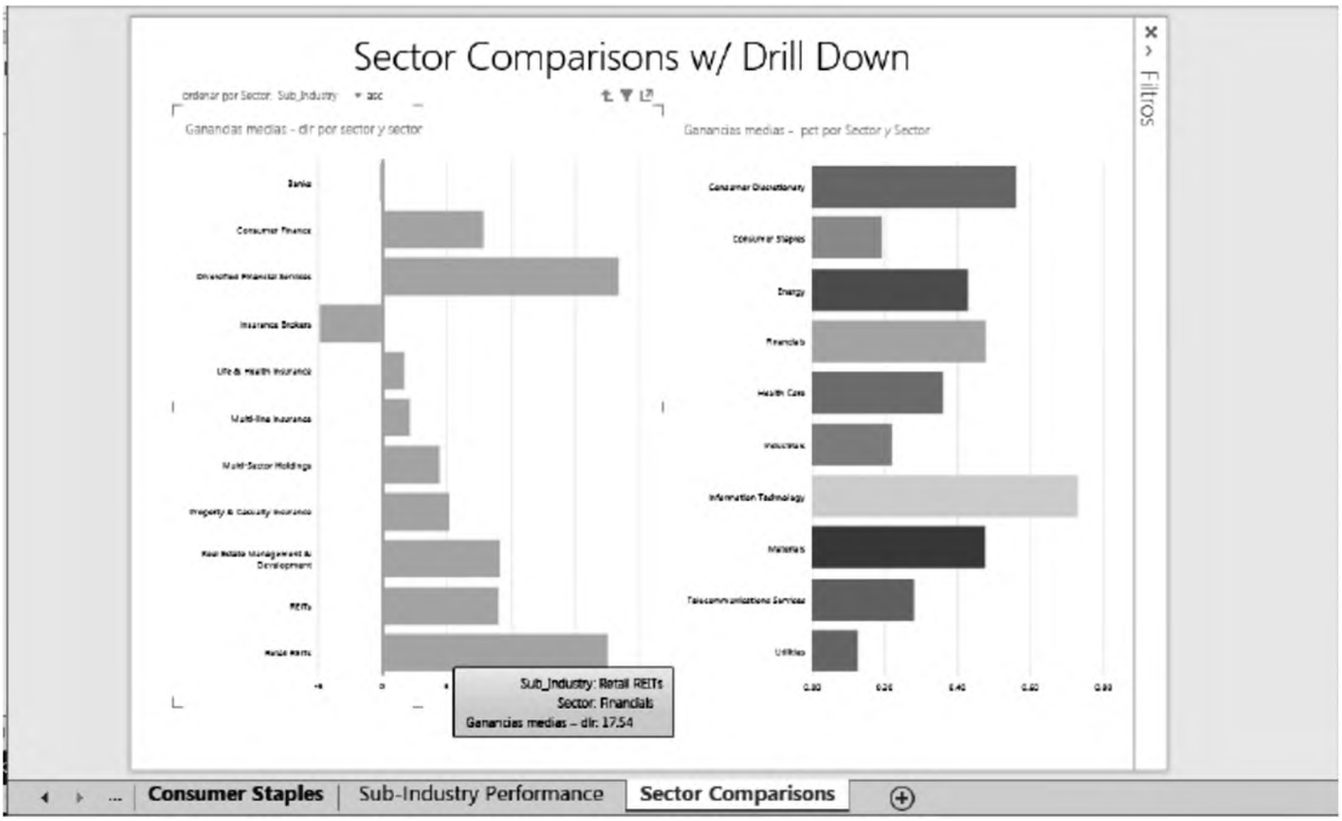


Figura 3-30

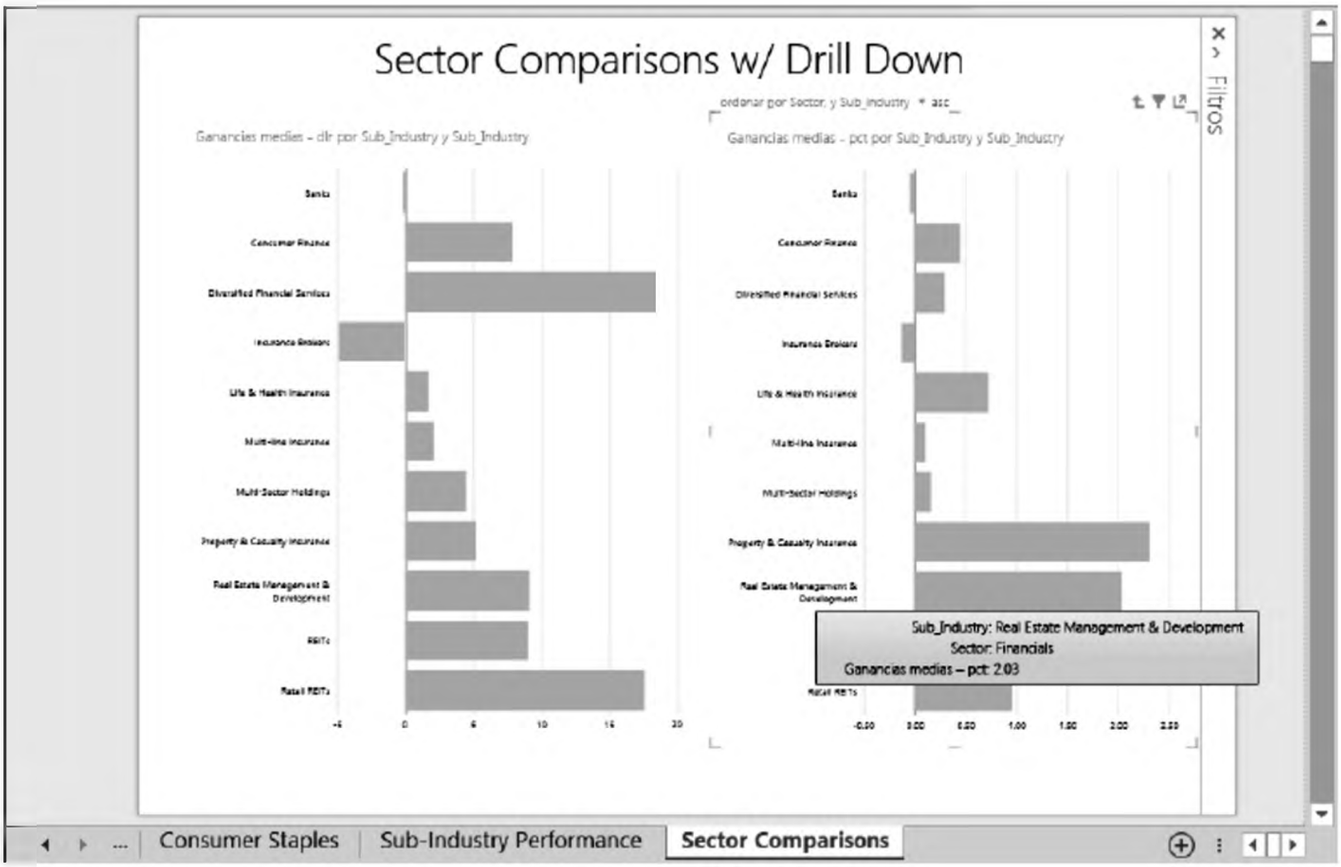


Figura 3-31

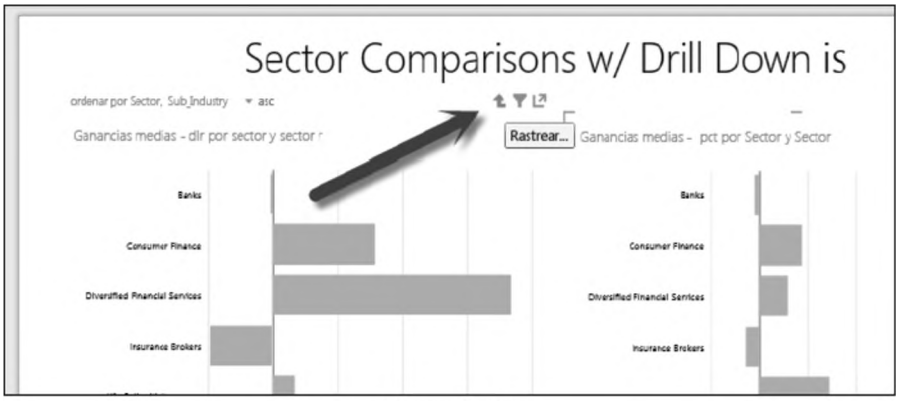


Figura 3-32

POWER MAP

Una de las últimas novedades de Excel es Power Map, una característica de Excel para la creación de mapas geocéntricos y temporales que convierte datos de ubicación en visualizaciones de mapas 3D eficaces e interesantes. Con Power Map, puede crear informes basados en la ubicación, que se visualizan a lo largo de una línea temporal y que permiten recorrer los datos visualmente y proporcionan perspectivas únicas.

Hace una búsqueda rápida en Bing y viendo lo que hacen otros usuarios con Power Map, está claro que merece la pena utilizar esta herramienta de visualización.

La Figura 3-33 muestra las centrales eléctricas de Estados Unidos, de 1900 a 2008.

La Figura 3-34 muestra el consumo de energía de un área cerca de Dallas, Texas, a lo largo del tiempo.

La Figura 3-35 muestra en un mapa las detenciones relacionadas con estupefacientes realizadas en Chicago.

En la Figura 3-36 se observa una visualización especializada de Power Map.

©Alfaomega-RC Libros

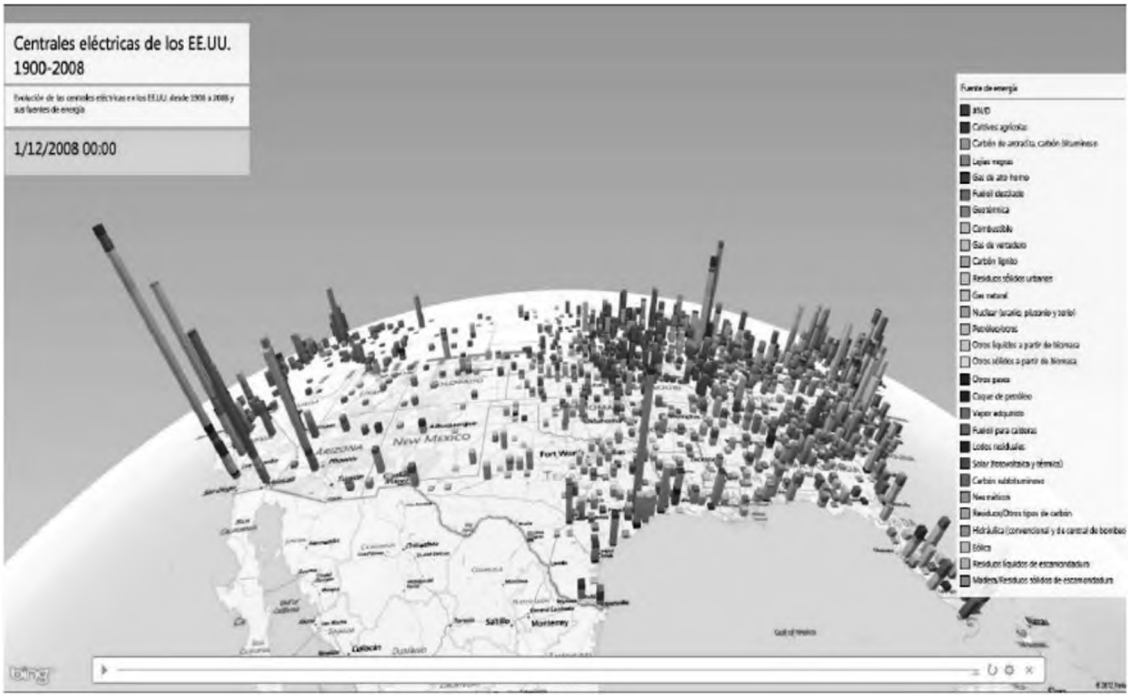


Figura 3-33

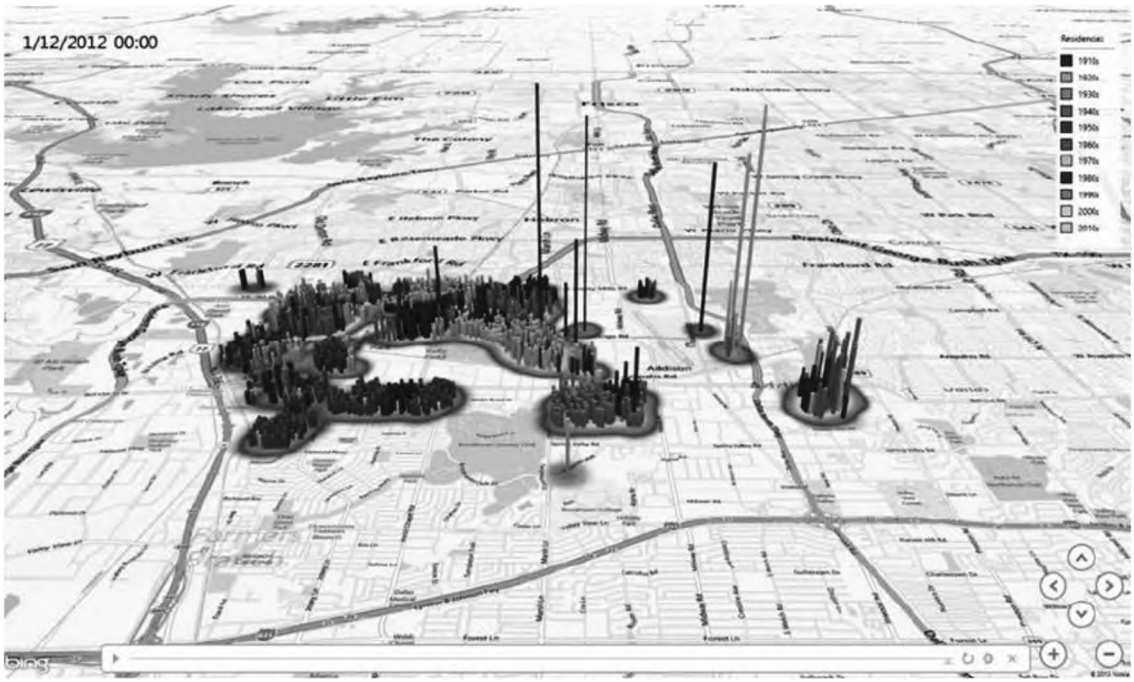


Figura 3-34

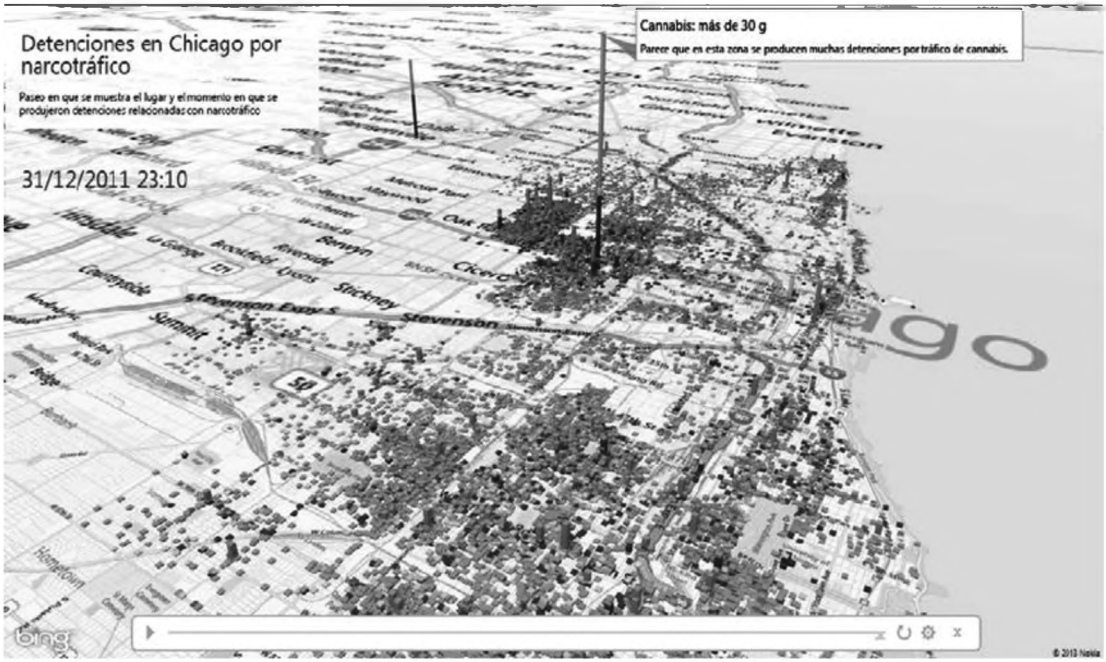


Figura 3-35

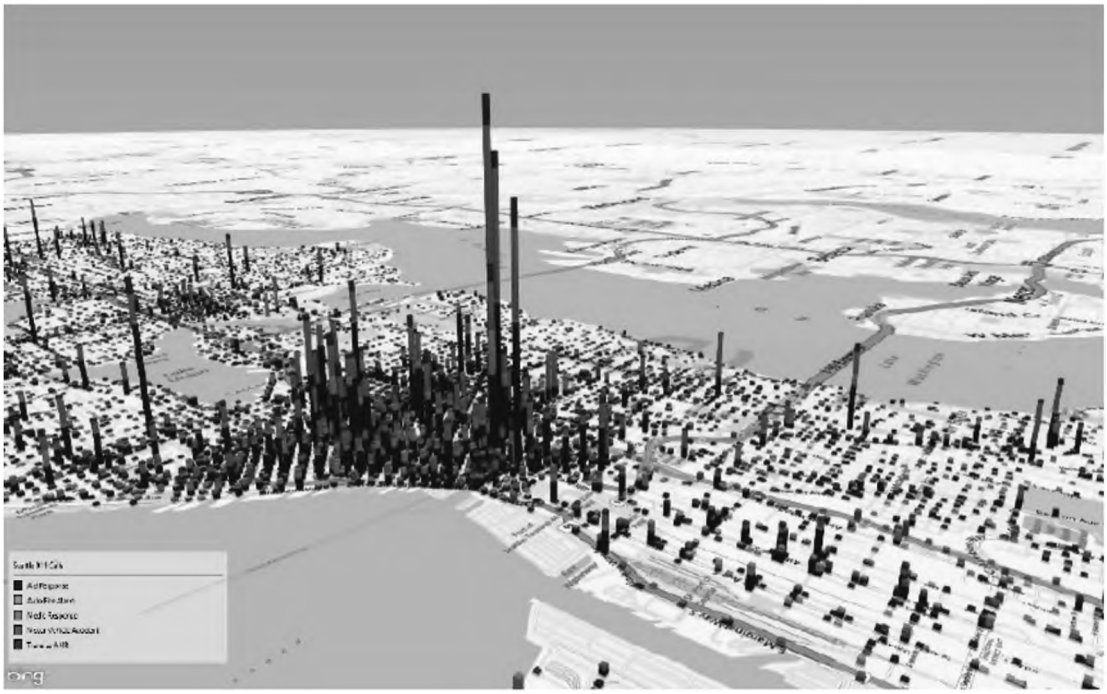


Figura 3-36

©Alfaomega-RC Libros

Con la característica de Excel Power Map, los mapas cobran un significado totalmente nuevo. Con la funcionalidad de modelado 3D y los paseos interactivos, Power Map puede hacer que los usuarios que visualizan los mapas vean los datos como en un vuelo. Así, lleva la interactividad a un nivel completamente nuevo.

Con los controles de zoom, desplazamiento lateral y altitud, Power Map es, aparte de los paseos interactivos, una forma impresionante de visualizar los datos. Seleccione puntos de datos, vea los datos desde el espacio o de cerca y elija cualquier punto de datos para obtener información. También puede crear notas que resalten la información directamente en el mapa y llevar los análisis y las perspectivas a un mundo visual en 3D.

CAPÍTULO 4

BUSINESS INTELLIGENCE CON MICROSOFT BI SQL SERVER

CARACTERÍSTICAS DE BUSINESS INTELLIGENCE EN MICROSOFT BI SQL SERVER

Las características de SQL Server que son parte de la plataforma Microsoft BI incluyen Analysis Services, Integration Services, Reporting Services y varias aplicaciones cliente que se usan para crear datos analíticos o para trabajar con ellos. Analysis Services y Reporting Services se pueden integrar con una granja de Microsoft SharePoint para habilitar las características de Business Intelligence (BI) en SharePoint. Entre las características se incluyen PowerPivot para SharePoint, Power View y Reporting Services. PowerPivot para SharePoint se usa para el acceso a datos PowerPivot en una granja de SharePoint. PowerPivot para SharePoint es el motor de datos para los libros creados en PowerPivot para Excel y accesibles desde una biblioteca de SharePoint. Una vez que guarde un libro de PowerPivot en SharePoint, puede utilizarlo como origen de datos para los informes de Power View.

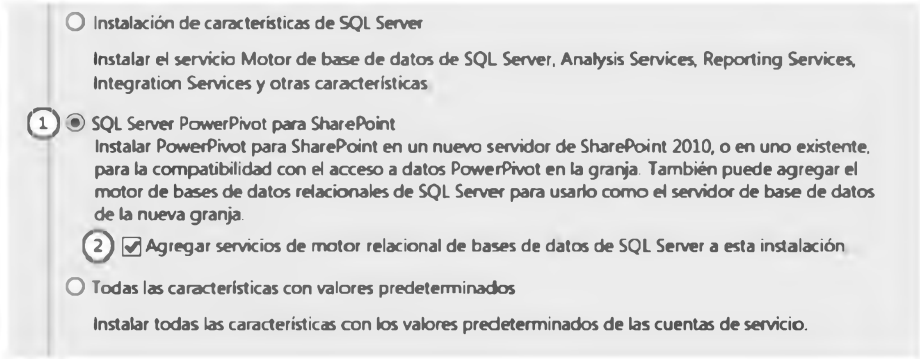
Si desea usar, tanto PowerPivot para SharePoint como Reporting Services, ejecute dos veces el Asistente para la instalación de SQL Server. Reporting Services y PowerPivot son opciones diferentes de la página Rol de instalación del Asistente para la instalación de SQL Server. PowerPivot para SharePoint admite SharePoint 2010 y SharePoint 2013; sin embargo, se utilizan una arquitectura y un proceso de instalación distintos en función de la versión de SharePoint.

A continuación se muestra un resumen de los pasos de instalación para implementar características de BI de SQL Server 2014 en un único servidor.

PowerPivot para SharePoint 2013

En **SharePoint 2013**, la instalación de PowerPivot para SharePoint se puede ejecutar en un servidor que no tiene productos SharePoint instalados. La arquitectura de PowerPivot que se usa para SharePoint 2013 se ejecuta **fuera** de la granja de servidores de SharePoint y se puede instalar en un servidor que también contiene una instalación de SharePoint o bien se puede instalar en un servidor que no contenga una instalación de SharePoint. Para instalar SQL Server PowerPivot:

1. Instale SharePoint Server 2013 y habilite Excel Services.
2. Instale Analysis Services en modo de SharePoint y conceda a las cuentas de servicios y de granja de SharePoint derechos de administrador del servidor en Analysis Services. En ambas versiones de SharePoint, el proceso de instalación de PowerPivot se inicia con la selección de la instalación del rol **SQL Server PowerPivot para SharePoint** en el Asistente para la instalación de SQL Server o utilizando una instalación desde el símbolo del sistema de SQL Server.



 ©Alfaomega-RC Libros

3. Para SharePoint 2013, puede ampliar las características y la experiencia de PowerPivot. Descargue y ejecute s**pPowerPivot.msi** para agregar compatibilidad con procesamiento, colaboración y administración de actualización de datos del servidor para los libros PowerPivot. Ejecute el paquete de instalación de PowerPivot para SharePoint 2013 **spPowerPivot.msi** en cada servidor de la granja de SharePoint para asegurarse de que se instala la versión correcta de los proveedores de datos.

4. Para configurar PowerPivot para SharePoint 2013, use la herramienta **Configuración de PowerPivot para SharePoint 2013**. El asistente para la instalación de SQL Server instala dos herramientas de configuración de PowerPivot. Una de las herramientas de configuración admite SharePoint 2013 y la otra admite SharePoint 2010.

5. Configure Excel Services en SharePoint Server 2013 para usar la instancia de Analysis Services.

PowerPivot para SharePoint es un servidor de Analysis Services en modo de SharePoint que proporciona hospedaje de servidor de datos en una granja de SharePoint. Los datos de PowerPivot son un modelo de datos analíticos que se genera mediante el Complemento PowerPivot para Excel 2010 o Excel 2013.

SQL Server 2014 PowerPivot para SharePoint admite el uso por parte de Excel Services de Microsoft SharePoint 2013 de libros de Excel que contienen modelos de datos e informes de Power View de Reporting Services.

Excel Services en SharePoint 2013 incluye funcionalidad de modelo de datos para habilitar la interacción con un libro PowerPivot en el explorador. No es necesario implementar el complemento PowerPivot para SharePoint 2013 en la granja de servidores. Solo se necesita instalar un servidor Analysis Services en modo de SharePoint y registrarlo dentro de la configuración de Modelo de datos de Excel Services.

La implementación del complemento PowerPivot para SharePoint 2013 habilita funcionalidad y características adicionales en la granja de servidores de SharePoint.

La Figura 4-1 muestra la implementación de servidores PowerPivot.

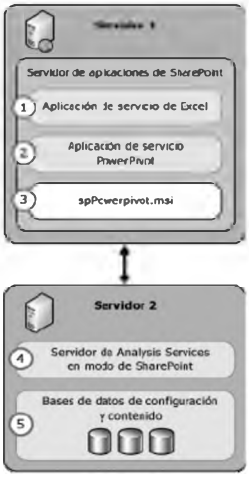


Figura 4-1

PowerPivot para SharePoint 2010

En SharePoint 2010, es necesario que la instalación de PowerPivot para SharePoint se ejecute en un servidor en el que se instalará SharePoint 2010 o que ya lo tenga instalado. La arquitectura de PowerPivot para SharePoint 2010 se ejecuta dentro de la granja y requiere SharePoint en el servidor en el que se ha instalado PowerPivot para SharePoint.

1. Instale Analysis Services en modo de SharePoint y conceda a las cuentas de servicios y de granja de SharePoint derechos de administrador del servidor en Analysis Services. Las implementaciones de SharePoint 2010 no admiten spPowerPivot.msi, y el archivo .msi no es necesario con SharePoint 2010. En ambas versiones de SharePoint, el proceso de instalación de PowerPivot se inicia con la selección de la instalación del rol SQL Server PowerPivot para SharePoint en el Asistente para la instalación de SQL Server o utilizando una instalación desde el símbolo del sistema de SQL Server.

2. El asistente para la instalación de SQL Server instala dos herramientas de configuración de PowerPivot. Una de las herramientas de configuración admite SharePoint 2013 y la otra admite SharePoint 2010. Para configurar PowerPivot para SharePoint 2010, use Herramienta de configuración de PowerPivot.

©Alfaomega-RC Libros

PowerPivot para SharePoint 2010 proporciona hospedaje de servidor de datos PowerPivot en una granja de SharePoint 2010. Los datos de PowerPivot son un modelo de datos analíticos creado en Excel mediante el complemento PowerPivot para Excel.

El hospedaje de servidor de esos datos requiere SharePoint 2010, Excel Services y una instalación de PowerPivot para SharePoint. Los datos se cargan en las instancias de PowerPivot para SharePoint en la granja, donde pueden actualizarse a intervalos programados mediante la función de actualización de datos PowerPivot que el servidor proporciona.

PowerPivot para SharePoint se implementa como un servicio compartido, lo que significa que las características integradas y la infraestructura se pueden utilizar para administrar, proteger y usar una aplicación de servicio PowerPivot. La detección, redirección y administración de conexiones del servidor y la base de datos se administran en el nivel de granja.

Administración central proporciona la interfaz administrativa a los servicios utilizados para administrar la identidad del servidor, el estado del servidor y las propiedades de configuración.

Una implementación completa de cliente de PowerPivot para SharePoint incluye componentes de servidor que se integran con Excel y Excel Services en una granja de servidores de SharePoint. Los datos PowerPivot de un libro de Excel son una base de datos de Analysis Services que requiere un motor analítico en memoria xVelocity de Analysis Services (VertiPaq) para cargar y consultar los datos. En una estación de trabajo cliente, el motor xVelocity se ejecuta en proceso dentro de Excel.

En una granja de servidores de SharePoint, Analysis Services se ejecuta en un servidor de aplicaciones donde está emparejado con servicios relacionados que administran solicitudes de datos PowerPivot.

El siguiente diagrama (Figura 4-2) ilustra los componentes de servidor y cliente de PowerPivot:

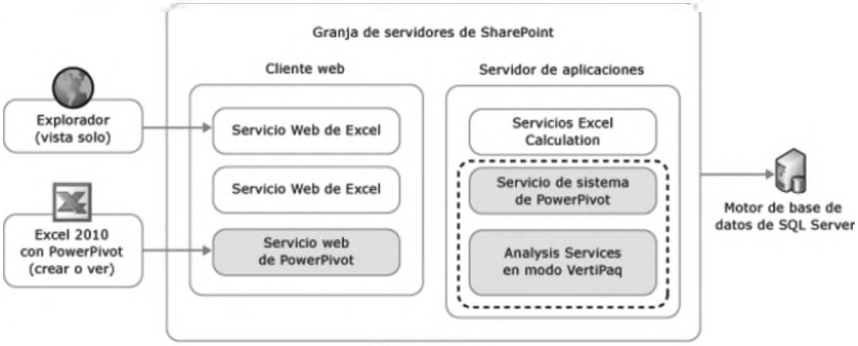


Figura 4-2

El servicio web de PowerPivot se ejecuta en un servidor de aplicaciones web. Redirige las solicitudes desde la aplicación Web a una instancia de Servicio de sistema de PowerPivot en la granja.

El Servicio de sistema de PowerPivot emite solicitudes de carga al servidor de Analysis Services y administra las conexiones salientes a los datos que ya están cargados en la memoria, almacenando en memoria caché o descargando los datos, si ya no se utilizan o si se produce alguna contención con los recursos del sistema. También realiza un seguimiento de la actividad de los usuarios. Los datos de estado de servidor y otros datos de uso se recopilan y se presentan en informes para indicar el grado de idoneidad del funcionamiento del sistema.

Una instancia de servidor de Analysis Services en modo integrado de SharePoint completa la implementación. Carga, consulta y descarga los datos. También procesa los datos si el libro se configura para la actualización de datos PowerPivot. Cada instancia está unida estrechamente al Servicio de sistema de PowerPivot local que forma parte de la misma instalación.

POWER VIEW PARA SHAREPOINT SERVER: CREAR, GUARDAR E IMPRIMIR INFORMES

Power View para SharePoint es una aplicación de Silverlight basada en explorador, una característica del complemento SQL Server Service Reporting Services para Microsoft SharePoint Server 2010 y 2013. Crear, abrir y guardar informes de Power View (archivos RDLX) son acciones que se realizan en SharePoint Server 2010 y 2013.

©Alfaomega-RC Libros

Crear un informe en Power View para SharePoint Server

Para crear un informe de Power View, inicie Power View desde un archivo de modelo de datos en SharePoint Server 2010 y 2013. Los modelos, o las conexiones a los modelos, pueden estar en una biblioteca de documentos de SharePoint Server o en una galería de PowerPivot, que es una biblioteca de documentos especial de SharePoint Server que permite obtener una completa vista previa y administrar los documentos de libros de Microsoft Excel publicados que contienen modelos de datos.

Para crear un informe de Power View desde una galería de PowerPivot:

Haga clic en el icono *Crear informe* de Power View en la esquina superior derecha del archivo de Excel (XLSX).

Para crear un informe de Power View desde una biblioteca de documentos de SharePoint Server:

Haga clic en un origen de datos compartido (RSDS) o archivo de conexión de BISM (BISM) para iniciar Power View.

El entorno de diseño de Power View se abre en la vista donde se crean los informes, con los campos del origen de datos compartido (RSDS) o el archivo de conexión de BISM (BISM) en la lista de campos.

Abrir un informe existente de Power View

Al abrir informes en una galería de PowerPivot, se puede optar por abrir el informe en una vista específica.

Para abrir un informe en una galería de PowerPivot:

- En Internet Explorer, vaya al sitio de galería de PowerPivot que hospeda los informes de Power View.
- Haga clic en cualquiera de las imágenes del informe.
- El informe se abre en la vista en modo de lectura.
- Para modificar el informe, haga clic en *Editar informe* en la esquina superior izquierda.

Para abrir un informe en una biblioteca de documentos de SharePoint Server

- En Internet Explorer, vaya a la página principal de la biblioteca de documentos de SharePoint que hospeda los informes de Power View.
- Para abrir un informe en modo de lectura, haga clic en el título del informe.
- Para modificar el informe, haga clic en *Editar informe* en la esquina superior izquierda.

Guardar un informe

Los informes de Power View (archivo RDLX) se guardan en una biblioteca de documentos compartidos o una galería de Power Pivot, en el mismo servidor de SharePoint que el modelo desde el que se ha iniciado Power View.

- Para guardar el informe, en el menú *Archivo* de Power View ssCrescent, haga clic en *Guardar* o en *Guardar como*. La primera vez que guarde el informe, la ubicación predeterminada será la carpeta donde se encuentra el modelo.
- Para guardarlo en una ubicación distinta, desplácese a esa ubicación y haga clic en *Guardar*. En el cuadro de diálogo *Guardar como*, en el campo *Nombre de archivo*, escriba el nombre del archivo.
- De forma predeterminada, la casilla *Guardar imágenes de vista previa con el informe* está seleccionada. Por razones de privacidad, quizá desee desactivarla y no guardar las imágenes de vista previa.
- Haga clic en *Guardar*.

El informe se guardará. Para salir de Power View y volver al sitio de SharePoint, haga clic en el botón *Atrás* del explorador.

Permisos para Power View

Power View usa permisos de SharePoint para controlar el acceso a los informes de Power View. Si tiene los permisos Abrir elementos para una carpeta de SharePoint, puede abrir cualquier informe de Power View de dicha carpeta en modo de edición o lectura. Por lo tanto, puede modificar el informe en modo de edición tanto como desee. Sin embargo, solo podrá guardar los cambios si dispone de los

©Alfaomega-RC Libros

permisos Agregar elementos en la biblioteca o carpeta de destino o los permisos Editar elementos para sobrescribir el documento existente.

También podrá exportar un informe a PowerPoint si dispone de los permisos Abrir elementos. Sin embargo, no podrá exportar informes a PowerPoint con cambios no guardados. Por lo tanto, si solo tiene permisos del tipo Abrir elementos, podrá exportar informes tal cual, pero no podrá modificarlos ni exportarlos. Para ello, primero deberá guardar los cambios, lo que implica que necesita los permisos Agregar elementos o Editar elementos.

Formato de archivo RDLX: Power View crea los archivos con el formato de archivo RDLX. Estos no son compatibles con los archivos RDL creados en el Generador de informes o SQL Server Reporting Services (SSRS). Los archivos RDL no se pueden abrir ni modificar en Power View, y viceversa.

Exportar a PowerPoint desde Power View en SharePoint

Puede exportar una versión interactiva del informe de Power View de SharePoint a PowerPoint. Cada vista de Power View se convierte en una diapositiva independiente de PowerPoint.

Hay dos versiones de Power View: Power View para Excel 2013 y Power View para SharePoint Server 2010 y 2013. Solo se puede exportar a PowerPoint desde Power View para SharePoint Server.

La interacción con los informes de Power View exportados a PowerPoint es similar a la interacción con las vistas de Power View en los modos de lectura y de pantalla completa de Power View. En los modos de presentación de diapositivas y de lectura de Microsoft PowerPoint, puede interactuar con las visualizaciones y los filtros que haya agregado el creador del informe a cada vista, pero no puede crear visualizaciones o filtros.

Actualizar los datos del informe

Puede actualizar los datos en un informe de Power View sin actualizar también la página. Haga clic en el botón *Actualizar* en la barra de herramientas de acceso rápido de Power View.

Si hace clic en el botón *Actualizar* en el explorador y, a continuación, hace clic en *Salir* de esta página, perderá los cambios realizados en el informe desde que lo guardó por última vez.

Imprimir vistas en un informe de Power View para SharePoint Server

Puede imprimir un informe de Power View en los modos de diseño o lectura, pero no en modo de pantalla completa. Power View imprime una vista a la vez: la vista actual.

1. Para imprimir una vista, en el menú *Archivo* de Power View haga clic en *Imprimir*. Se abre el cuadro de diálogo *Imprimir* del explorador.
2. Haga clic en *Imprimir*. La vista siempre se imprime en orientación horizontal, independientemente de la configuración del cuadro de diálogo *Imprimir*. Imprime exactamente lo que aparece en la vista. Por ejemplo, Power View imprime:

    - La parte de una visualización que está visible al imprimir, si la visualización tiene una barra de desplazamiento.
    - El mosaico seleccionado en un contenedor en mosaico.
    - El área de filtros, si se expande.
    - El marco actual de un gráfico de dispersión o burbujas con un eje de reproducción.

Informes de Power View basados en modelos de datos

En SharePoint, Power View se inicia siempre desde un modelo de datos. El modelo puede ser:

1. Un libro de Excel (XLSX) con un modelo de datos en una galería de PowerPivot en SharePoint Server. Usar una galería de PowerPivot.
2. Un origen de datos compartidos (RSDS) en SharePoint Server con un tipo de origen de datos de Microsoft Business Intelligence Semantic Model, basado en:

    - Un libro de Excel.
    - Un modelo tabular en un servidor de Analysis Services. Crear un origen de datos compartido para un modelo de datos.

©Alfaomega-RC Libros

- Un modelo multidimensional en un servidor SSAS. Descripción de los objetos de modelo multidimensional de Power View.
- Un archivo de conexión de BISM (BISM) basado en un modelo tabular en un servidor de Analysis Services. Los archivos de conexión de BISM pueden estar en una biblioteca de documentos estándar de SharePoint Server o en una galería de PowerPivot. Usar una conexión de modelo semántico de BI.

Los modelos de datos son la siguiente generación de modelos diseñada específicamente para las cargas de trabajo de informes y análisis. Los modelos de datos pueden importar datos desde una gran variedad de orígenes de datos, entre los que se incluyen:

- SQL Server
- DB2
- OData
- Oracle
- Teradata
- Y otros

Se puede crear un modelo de datos en Excel, mejorar el modelo de datos con PowerPivot para Excel y después guardar el modelo en una biblioteca de documentos de SharePoint Server 2013 o en una galería de PowerPivot. Los programadores de modelos de una organización de TI crean modelos en SQL Server Data Tools (SSDT) y, después, los implementan en un servidor de SQL Server Analysis Services (SSAS).

Descripción de los objetos de modelo multidimensional de Power View

Se puede usar Power View iniciado desde SharePoint Server para explorar datos de manera interactiva y crear visualizaciones dinámicas de modelos multidimensionales de Analysis Services en Microsoft SQL Server.

Cuando se usa Power View para visualizar modelos multidimensionales, es importante que tenga en cuenta que está trabajando con una representación de tipo de modelo tabular de un modelo multidimensional. Los modelos tabulares tienen objetos como tablas y columnas, y al igual que con los modelos multidimensionales, medidas y KPI.

Cuando usted o un administrador crea una conexión de origen de datos compartida en SharePoint, se especifica un nombre o perspectiva de cubo nombre en la cadena de conexión. Solo puede especificarse un cubo o perspectiva. El cubo o la perspectiva especificada en la conexión a un origen de datos compartida se exponen como modelo en la Lista de campos de Power View. Los objetos del modelo se exponen como campos que se pueden usar como visualizaciones de tabla en una vista. Hay, sin embargo, algunas diferencias en cómo aparecen ciertos objetos multidimensionales en Power View. Al igual que con los modelos tabulares, la Lista de campos muestra todos los objetos que se pueden usar en una vista.

Los modelos multidimensionales tienen dimensiones. En este ejemplo, la Lista de campos contiene objetos de la dimensión Producto (Figura 4-3). El producto aparece como una tabla en el modelo Adventure Works (cubo). Una tabla o dimensión tiene también otros objetos.

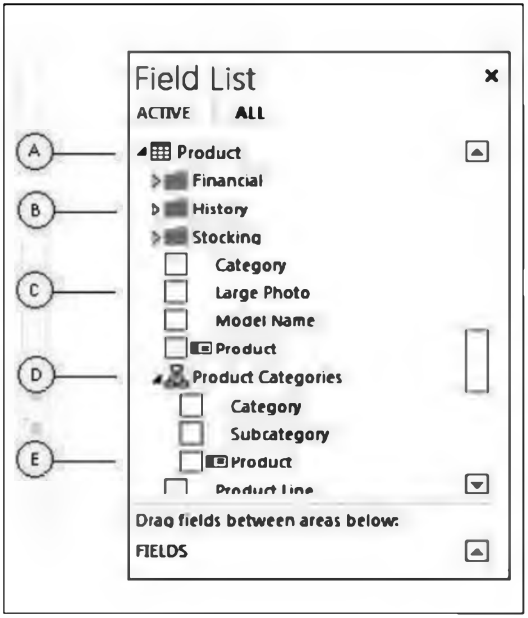


Figura 4-3

a) Las dimensiones aparecen como tablas y pueden expandirse para mostrar otros objetos en la tabla (dimensión). El cubo de Adventure Works tiene muchas tablas, cuenta a través de moneda de origen.

©Alfaomega-RC Libros

b) Mostrar carpetas aún más, dividir y clasificar cómo aparecen los objetos asociados en las herramientas de cliente. Mostrar carpetas puede aparecer en la Lista de campos debajo de las tablas de dimensiones y de las tablas de grupos de medida.

c) Los atributos de dimensión aparecen como columnas en una tabla. Una única columna (atributo) puede aparecer en una tabla y de nuevo en una jerarquía, a menos que se oculte explícitamente.

d) Las jerarquías de usuario y elementos primarios y secundarios se incluyen en las tablas (dimensiones). Las jerarquías se pueden expandir para mostrar columnas (niveles) en ellas. Cuando se selecciona un nivel, todos los niveles anteriores también se seleccionan automáticamente. Puede deseleccionar los niveles más altos para quitarlos de la visualización. Esto puede resultar útil cuando determinados campos se exponen solo en una jerarquía.

e) Los objetos con este icono indican que objeto es una clave.

Los modelos multidimensionales también tienen grupos de medidas (Figura 4-4), también conocidos como dimensiones de medidas, que contienen las medidas que se pueden utilizar para agregar datos en el análisis.

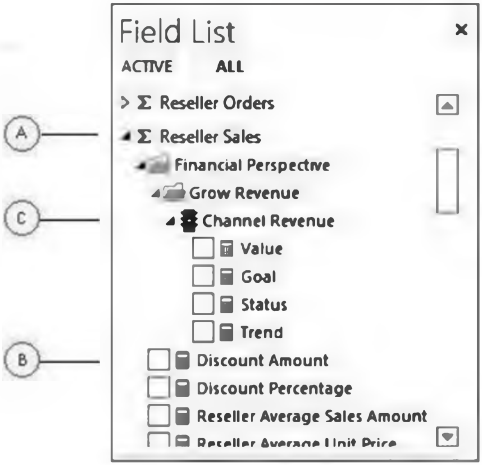


Figura 4-4

a) Los grupos de medidas aparecen como tablas; sin embargo, a diferencia de las tablas de una dimensión, una tabla de un grupo de medidas se identifica con un icono sigma.

b) Las medidas aparecen en tablas de grupos de medidas y se identifican con un icono de calculadora. Si el cubo tiene solo una medida, se incluirá en un grupo de medidas asociado si hay uno, o en una sola tabla denominada Medidas.

c) Los KPI se incluyen en tablas de grupos de medidas asociadas y se identifican con un icono de luz. Por ejemplo, si en el modelo de Adventure Works expande Venta del distribuidor > Perspectivas financieras > Aumentar los ingresos, verá el KPI de Ingresos de canal y sus cuatro medidas: Valor, Objetivos, Estado y Tendencia.

Gráficos y otras visualizaciones en Power View

En Power View en SharePoint 2013 y Excel 2013, puede crear rápidamente una variedad de visualizaciones de datos, desde tablas y matrices hasta gráficos de barras, columnas y burbujas, así como conjuntos de gráficos de múltiplos. Sea cual sea la visualización que desee crear, siempre comenzará en una hoja de Power View creando una tabla, que más adelante se convertirá fácilmente en otras visualizaciones para determinar cuál es la que mejor ilustra los datos.

Para crear una visualización:

1. Cree una tabla en la hoja de Power View activando una tabla o campo en la lista de campos o arrastrando un campo de la lista de campos a la hoja. Power View dibuja la tabla en la hoja, muestra los datos reales y agrega automáticamente encabezados de columna.

2. Convierta la tabla en una visualización eligiendo un tipo de visualización en la pestaña Diseño. Dependiendo de los datos de la tabla, diferentes tipos de visualizaciones estarán disponibles para darle la mejor visualización para esos datos.

A continuación se presenta una visualización típica en Power View (Figura 4-5):

©Alfaomega-RC Libros

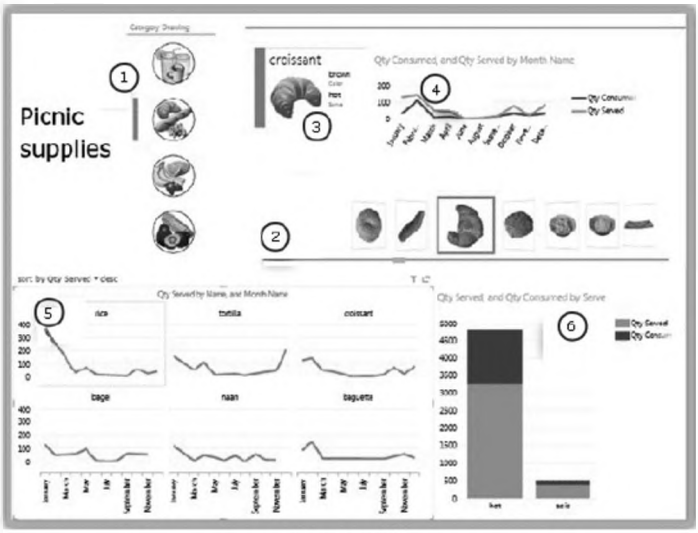


Figura 4-5

Sus elementos son los siguientes:

1. Segmentación que filtra el informe por panes.

2. Navegación de flujo de mosaicos para los mosaicos, actualmente en croissant.

3. Tarjeta en un contenedor en mosaico, filtrado por el mosaico actual (croissant).

4. Gráfico de líneas en el contenedor en mosaico que muestra las cantidades consumidas y servidas, filtrado por croissants de enero a diciembre.

5. Múltiplos, filtrados por panes y ordenados en orden descendente por cantidad servida.

6. Gráfico de columnas filtrado por panes, que muestra las cantidades servidas y consumidas.

Todas las visualizaciones comienzan con una tabla (Figura 4-6).

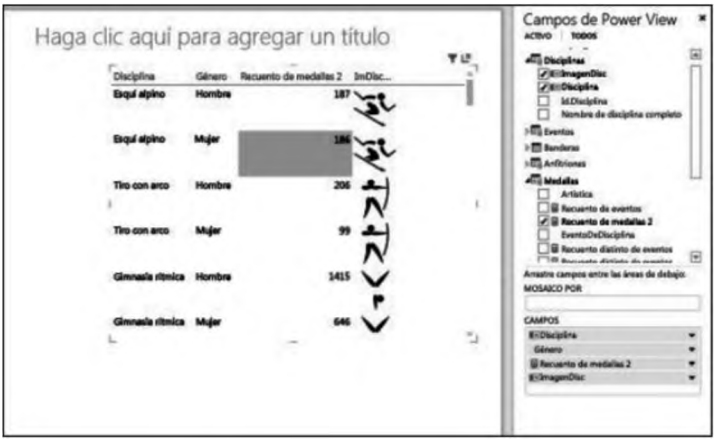


Figura 4-6

Power View ofrece una serie de opciones de gráficos: circulares, columnas, barras, líneas, dispersión y burbujas. Los gráficos pueden tener varios campos numéricos y varias series. En cuanto a las opciones de diseño en un gráfico puede mostrar u ocultar etiquetas, leyendas y títulos.

Los gráficos son interactivos. Al hacer clic en un valor de un gráfico, se:

- Resalta ese valor en el gráfico.
- Filtra por ese valor en todas las tablas, matrices y mosaicos del informe.
- Destaca ese valor en todos los demás gráficos del informe.

Los gráficos también son interactivos en las presentaciones. Por ejemplo, en los modos de lectura y de pantalla completa de Power View en SharePoint o de una hoja de Power View en un libro de Excel guardado en Excel Services o visto en Office 365.

Gráficos circulares

Los gráficos circulares de Power View pueden ser simples o sofisticados. Puede crear un gráfico circular que desglose la información al hacer doble clic en un sector o un gráfico circular que muestre subsectores dentro de los sectores de color más grandes (Figura 4-7).

©Alfaomega-RC Libros

Puede aplicar filtros cruzados de un gráfico circular a otro gráfico. Imagine que hace clic en una barra de un gráfico de barras: la parte del gráfico circular que se aplica a esa barra se resalta y se atenúa el color del resto del gráfico.

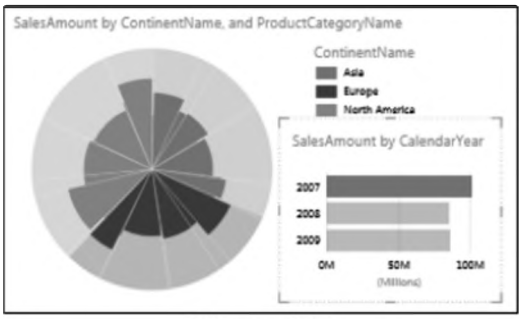


Figura 4-7

Gráficos de dispersión y de burbujas

Los gráficos de dispersión y de burbujas son una magnífica forma de mostrar grandes cantidades de datos relacionados en un mismo gráfico. En los gráficos de dispersión (Figura 4-8), el eje X muestra un campo numérico y el eje Y muestra otro, de modo que resulta fácil ver la relación entre los dos valores para todos los elementos del gráfico.

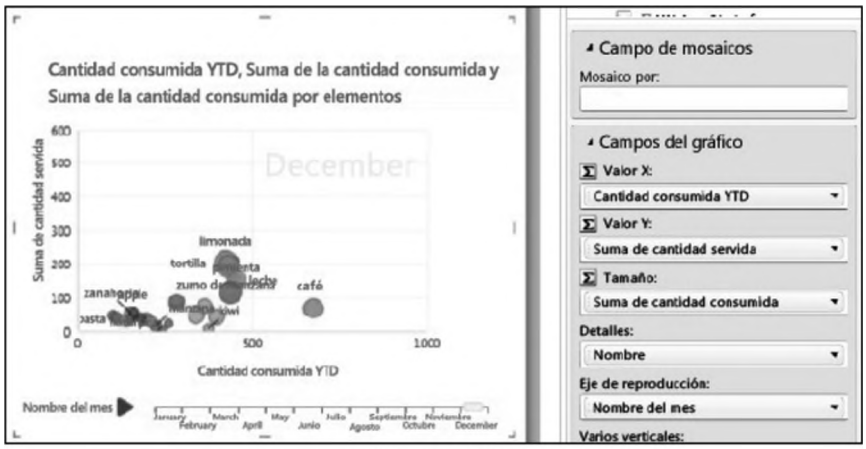


Figura 4-8

En un gráfico de burbujas (Figura 4-9), un tercer campo numérico controla el tamaño de los puntos de datos. Puede agregar un eje de "reproducción" a un gráfico de burbujas o dispersión para ver los datos conforme cambian con el tiempo.

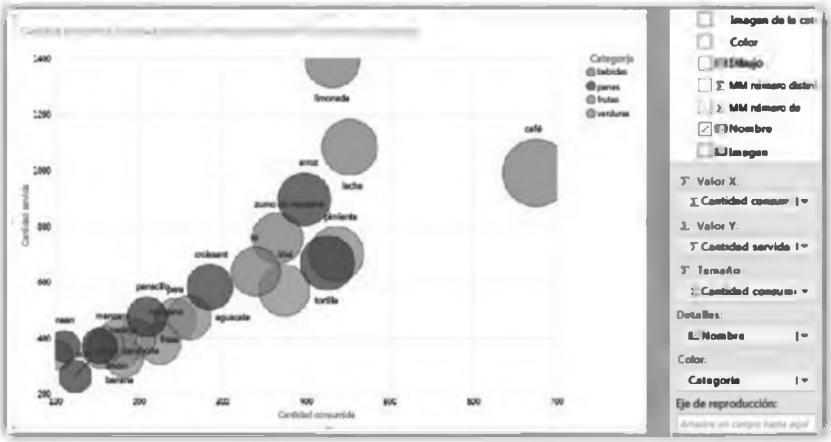


Figura 4-9

Gráficos de líneas, barras y columnas

Los gráficos de líneas, barras y columnas son útiles para comparar puntos de datos en una o varias series de datos. En los gráficos de líneas, barras y columnas, el eje X muestra un campo y el eje Y otro. De este modo es fácil ver la relación entre los dos valores de todos los elementos del gráfico.

Gráficos de barras

En los gráficos de barras, las categorías se organizan en torno al eje vertical y los valores en torno al eje horizontal. Considere la posibilidad de usar un gráfico de barras en cualquiera de estos casos:

- Tiene una o más series de datos que desea representar.
- Sus datos incluyen valores positivos, negativos y cero (0).
- Desea comparar los datos de varias categorías.
- Las etiquetas de eje son largas.

En Power View, dispone de tres subtipos de gráficos de barras para elegir: apilados, 100% apilados y agrupados.

©Alfaomega-RC Libros

Gráficos de columnas

Los datos organizados en columnas o filas en una hoja de cálculo se pueden representar en un gráfico de columnas. Los gráficos de columnas resultan útiles para mostrar cambios en los datos a lo largo de un período de tiempo o para ilustrar comparaciones entre elementos. En los gráficos de columnas, las categorías se organizan en torno al eje horizontal y los valores se organizan en torno al eje vertical.

En Power View, dispone de tres subtipos de gráficos de columnas para elegir: apilados, 100% apilados y agrupados.

Gráficos de líneas

Los gráficos de líneas (Figura 4-10) distribuyen los datos de categorías de manera uniforme en un eje horizontal (categoría) y distribuyen todos los datos de valores numéricos a lo largo de un eje vertical (valor).

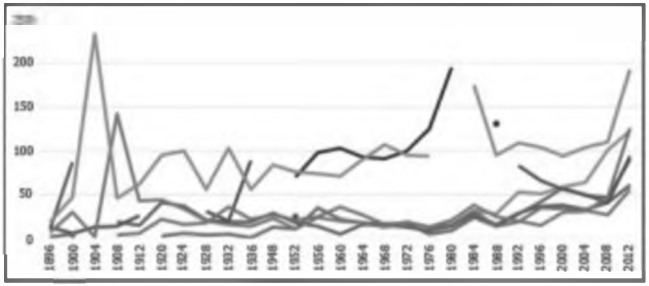


Figura 4-10

Considere la posibilidad de usar un gráfico de líneas con escala de tiempo en el eje horizontal. Los gráficos de líneas muestran las fechas en orden cronológico a intervalos concretos o unidades básicas, como el número de días, meses o años, incluso si las fechas de la hoja de cálculo no están en orden o en las mismas unidades básicas.

Mapas

Los mapas de Power View (Figura 4-11) utilizan los mosaicos de los mapas de Bing, por lo que puede hacer zoom y desplazarse como haría en cualquier otro mapa de Bing. Al agregar lugares y campos, se colocan puntos en el mapa. Cuanto mayor sea el valor, mayor será el punto. Cuando se agrega una serie de valores múltiples, se colocan

gráficos circulares en el mapa y el tamaño del gráfico circular muestra el tamaño del total. Encontrará más información sobre los mapas en Power View.

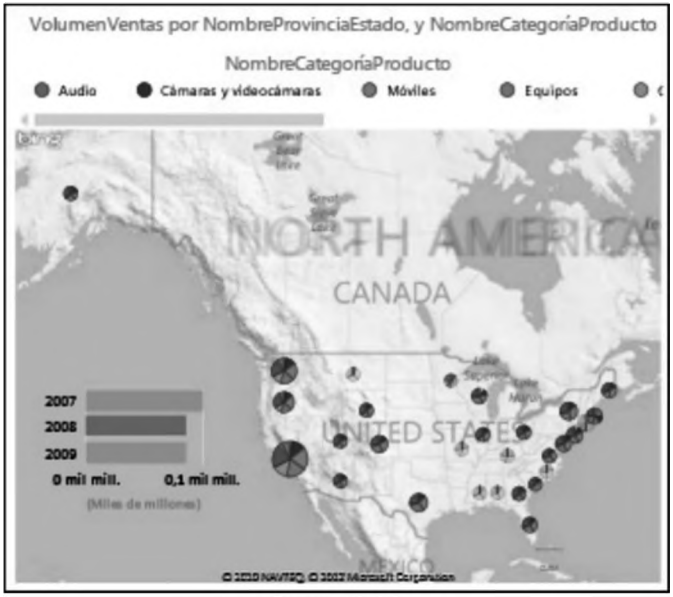


Figura 4-11

Múltiplos: un conjunto de gráficos con los mismos ejes

Con los múltiplos, puede crear una serie de gráficos con ejes X e Y idénticos para, a continuación, organizarlos unos junto a otros, lo que facilita la comparación de distintos valores al mismo tiempo. Los múltiplos también suelen llamarse "gráficos enrejados" (Figura 4-12).

Figura 4-12

©Alfaomega-RC Libros

Matrices

Una matriz es similar a una tabla por el hecho de que está compuesta por filas y columnas. Sin embargo, las matrices tienen unas capacidades que no tienen las tablas:

- Muestran los datos sin repetir los valores.
- Muestran los totales y los subtotales por fila y columna.
- Con una jerarquía, es posible agregar detalles y resúmenes.
- Es posible expandir y contraer la visualización.

Tarjetas

Puede convertir una tabla en una serie de tarjetas que muestren los datos de cada fila de la tabla diseñada en formato de tarjeta, como una tarjeta de índice (Figura 4-13).

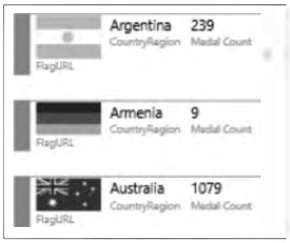


Figura 4-13

Mosaicos

Puede convertir una tabla o matriz en mosaicos para presentar los datos tabulares interactivamente. Los mosaicos son contenedores con una franja de navegación dinámica. Los mosaicos sirven de filtros: filtran el contenido que hay en el mosaico en función del valor seleccionado en la franja de pestañas. Puede agregar más de una visualización al mosaico y filtrar todas las visualizaciones por el mismo valor. Puede usar texto o imágenes como pestañas. Más información acerca de los mosaicos en Power View.

Esta imagen (Figura 4-14) muestra el número de medallas que han ganado los distintos países en una competición de patinaje de velocidad.

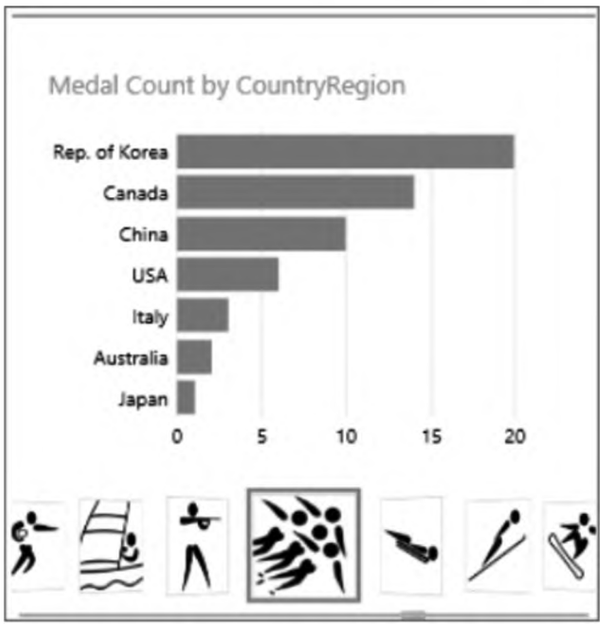
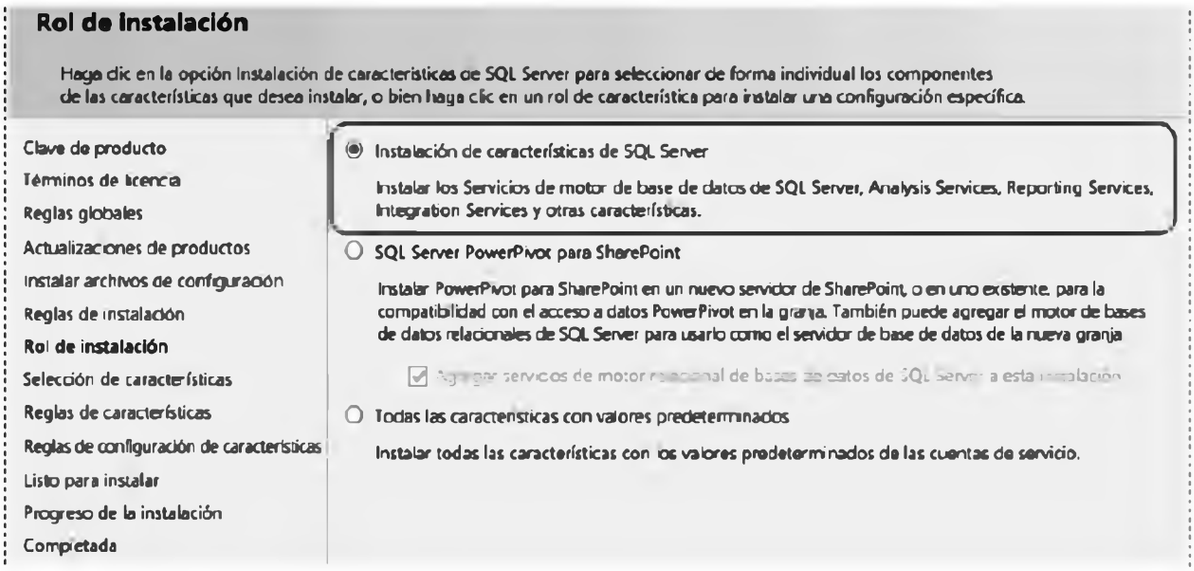


Figura 4-14

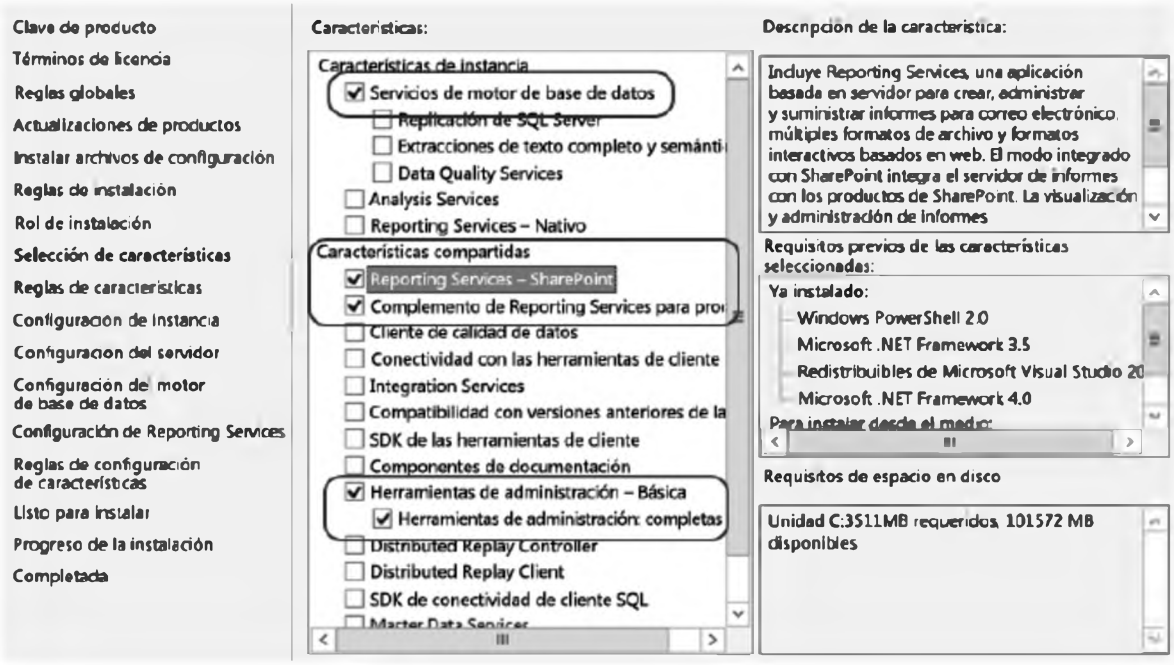
ANALYSIS SERVICES, INTEGRATION SERVICES Y REPORTING SERVICES

Para instalar Analysis Services, Integration Services y Reporting Services, basta elegir estas características BI de SQL Server, en la página *Rol de instalación* del Asistente para la instalación de SQL Server:



Para instalar Complemento de Reporting Services para productos de SharePoint, basta elegirlo en la página Selección de características del Asistente para la instalación de SQL Server:

©Alfaomega-RC Libros



Informes de Reporting Services (SSRS)

Los informes de SQL Server Reporting Services son definiciones de informe basadas en XML que incluyen los datos y elementos de diseño de los informes (Figura 4-15). En un sistema de archivos de cliente, las definiciones de informe tienen la extensión de archivo .rdl. Una vez que se publica un informe, se convierte en un elemento de informe que se almacena en el servidor de informes o en el sitio de SharePoint. Los informes constituyen un único componente de la plataforma de generación de informes basada en servidor que proporciona Reporting Services.

Puede usar las soluciones de informes de Reporting Services para:

- Usar un solo conjunto de orígenes de datos que proporcione una única versión de los hechos. Puede basar los informes en esos orígenes de datos para proporcionar una vista de datos unificada que facilite la toma de decisiones comerciales.

- Visualizar los datos de formas diversas e interconectadas a través de las regiones de datos. Puede mostrar los datos organizados en tablas, matrices o tablas de referencias cruzadas; también puede expandir o contraer grupos, gráficos, medidores, indicadores o KPI, y mapas, e incluso tiene la posibilidad de incluir gráficos en las tablas.

- Ver los informes para su propio uso o publicar informes en un servidor de informes o un sitio de SharePoint para compartirlos con el equipo o la organización.

- Definir un informe una sola vez y presentarlo de diversas maneras. Puede exportar el informe en varios formatos de archivo o entregar el informe a los suscriptores en forma de correo electrónico o de un archivo compartido. Puede crear informes vinculados que apliquen distintos conjuntos de parámetros a la misma definición de informe.

- Usar elementos de informe, orígenes de datos compartidos, consultas compartidas y subinformes para definir las visualizaciones de datos para su reutilización.

- Administrar los orígenes de datos del informe con independencia de la definición de informe. Por ejemplo, puede cambiar de un origen de datos de prueba a un origen de datos de producción sin cambiar el informe.

- Crear informes con un diseño libre. El diseño del informe no está restringido a bandas de información. Puede organizar la visualización de los datos de la página de forma que facilite su comprensión, mejore su entendimiento y promueva la entrada en acción.

- Habilitar acciones para la obtención de detalles, alternadores para expandir y contraer, botones de ordenación, información sobre herramientas y parámetros de informe que permitan al lector interactuar con el informe. Puede combinar los parámetros de informe con sus propias expresiones para que los lectores del informe puedan controlar el modo en que se filtran, agrupan y ordenan los datos.

- Definir expresiones que le proporcionan la capacidad de personalizar el modo en que se filtran, agrupan y ordenan los datos.

Al crear un informe, tiene que definir un archivo de definición de informe (.rdl) en formato XML. Este archivo contiene toda la información necesaria para combinar los datos y el diseño del informe mediante el procesador de informes. Cuando consulte un informe, este avanzará a través de los pasos siguientes:

©Alfaomega-RC Libros

- Compilación. Se evalúan las expresiones de la definición de informe y se almacena el formato intermedio de compilación internamente en el servidor de informes.

- Proceso. Se ejecutan las consultas de conjuntos de datos y se combina el formato intermedio con los datos y el diseño.

- Representación. El informe procesado se envía a una extensión de representación para determinar cuánta información cabe en cada página y crear el informe paginado.

- Exportación (opcional). El informe se exporta a un formato de archivo diferente.

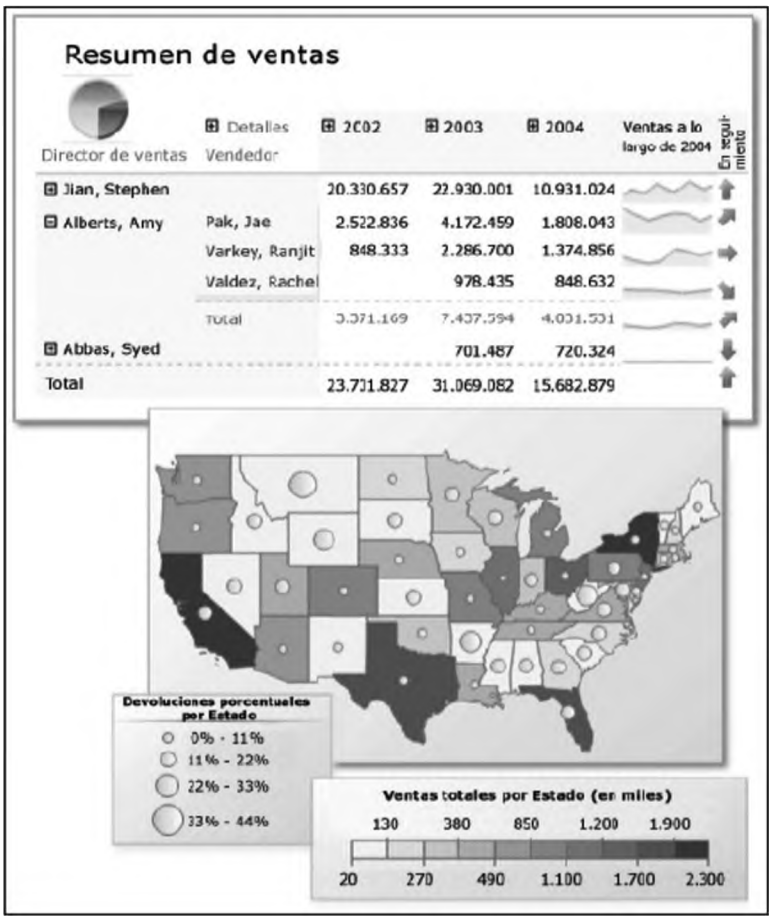


Figura 4-15

SQL Server Integration Services

Microsoft Integration Services es una plataforma para la creación de soluciones empresariales de transformaciones de datos e integración de datos. Integration Services sirve para resolver complejos problemas empresariales mediante la copia o descarga de archivos, el envío de mensajes de correo electrónico como respuesta a eventos, la actualización de almacenamientos de datos, la limpieza y minería de datos, y la administración de objetos y datos de SQL Server. Los paquetes pueden funcionar en solitario o junto con otros paquetes para hacer frente a las complejas necesidades de la empresa. Integration Services puede extraer y transformar datos de diversos orígenes como archivos de datos XML, archivos planos y orígenes de datos relacionales y, después, cargar los datos en uno o varios destinos.

Integration Services contiene un variado conjunto de tareas y transformaciones integradas, herramientas para la creación de paquetes y el servicio Integration Services para ejecutar y administrar los paquetes. Las herramientas gráficas de Integration Services se pueden usar para crear soluciones sin escribir una sola línea de código. También se puede programar el amplio modelo de objetos de Integration Services para crear paquetes mediante programación y codificar tareas personalizadas y otros objetos de paquete.

Un paquete es una colección organizada de conexiones, elementos de flujo de control, elementos de flujo de datos, controladores de eventos, variables, parámetros y configuraciones que se pueden ensamblar con la ayuda de las herramientas gráficas de diseño proporcionadas por SQL Server Integration Services o mediante programación. A continuación guarda el paquete completado en SQL Server, el Almacén de paquetes SSIS o el sistema de archivos, o puede implementar el proyecto de SSnoversion en el servidor SSIS. El paquete es la unidad de trabajo que se recupera, ejecuta y guarda.

Al crear por primera vez un paquete, es un objeto vacío que no hace nada. Para agregar funcionalidad a un paquete, debe agregarle un flujo de control y, opcionalmente, uno o más flujos de datos.

El siguiente diagrama (Figura 4-16) muestra un paquete individual que contiene un flujo de control con una tarea Flujo de datos que, a su vez, contiene un flujo de datos.

©Alfaomega-RC Libros

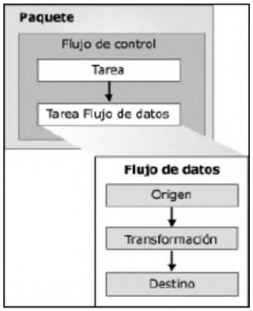


Figura 4-16

Una vez creado el paquete básico, puede agregarle características avanzadas como registro y variables para extender su funcionalidad.

Una vez completado el paquete, puede configurarse estableciendo propiedades de nivel de paquete que permitan implementar la seguridad, reiniciar paquetes desde puntos de comprobación o incorporar transacciones en el flujo de trabajo de paquetes.

Un flujo de control consta de una o más tareas y contenedores que se ejecutan cuando se ejecuta el paquete. Para controlar el orden o definir las condiciones para ejecutar la siguiente tarea o contenedor del flujo de control del paquete, puede usar restricciones de precedencia para conectar las tareas y los contenedores de un paquete. También se puede agrupar y ejecutar repetidamente un subconjunto de tareas y contenedores como una unidad en el flujo de control del paquete.

Un flujo de datos consta de los orígenes y destinos que extraen y cargan datos, las transformaciones que modifican y extienden datos, y las rutas que vinculan orígenes, transformaciones y destinos. Para poder agregar un flujo de datos a un paquete, el flujo de control de paquetes debe incluir una tarea Flujo de datos. La tarea Flujo de datos es el ejecutable del paquete SSIS que crea, organiza y ejecuta el flujo de datos. Se abre una instancia independiente del motor de flujo de datos para cada tarea Flujo de datos de un paquete.

Un paquete suele incluir al menos un administrador de conexiones. Un administrador de conexiones es un vínculo entre un paquete y un origen de datos que define la cadena de conexión para acceder a los datos que las tareas, transformaciones y controladores de eventos del paquete usan. Integration Services incluye topos de conexiones para orígenes de datos tales como archivos de texto y XML, bases de datos relacionales y proyectos y bases de datos de Analysis Services.

Los paquetes se utilizan con frecuencia como plantillas para generar paquetes que comparten una funcionalidad básica. Puede generar el paquete básico y luego copiarlo, o puede designar que el paquete sea una plantilla. Por ejemplo, un paquete que descarga y copia archivos, y luego extrae los datos, puede incluir las tareas FTP y Sistema de archivos en un bucle Foreach que enumera archivos de una carpeta. También podría incluir administradores de conexión de archivos planos para el acceso a los datos, y orígenes de archivos planos para la extracción de los datos. El destino de los datos varía y se agrega a cada nuevo paquete una vez copiado del paquete básico. También puede crear paquetes y luego usarlos como plantillas para los nuevos paquetes que agregue a un proyecto de Integration Services.

Analysis Services

Analysis Services es un motor de datos analíticos en línea que se usa en soluciones de ayuda a la toma de decisiones y Business Intelligence (BI), y proporciona los datos analíticos para informes empresariales y aplicaciones cliente como Excel, informes de Reporting Services y otras herramientas de BI de terceros. Un flujo de trabajo típico para Analysis Services incluye la creación de un modelo de datos OLAP o tabular, la implementación del modelo como base de datos en una instancia de Analysis Services, el procesamiento de la base de datos para cargarla con datos y, a continuación, la asignación de permisos para permitir el acceso a datos. Cuando esté listo, se puede obtener acceso a este modelo de datos con varios fines desde cualquier aplicación cliente que admita Analysis Services como origen de datos.

Para crear un modelo, use SQL Server Data Tools y elija una plantilla de proyecto Tabular o Multidimensional y Minería de datos. Los modelos se rellenan con datos procedentes de sistemas de datos externos, normalmente almacenamientos de datos hospedados en un motor de base de datos relacional de SQL Server o de Oracle (los modelos tabulares admiten tipos de orígenes de datos adicionales). Los modelos especifican objetos de consulta, como cubos y dimensiones, cálculos y KPI, además de interacciones como los comportamientos de navegación y de obtención de detalles.

Para usar un modelo, se implementa en una instancia de Analysis Services que ejecuta bases de datos en un modo de servidor determinado, haciendo que los datos estén disponibles para los usuarios autorizados que se conectan a través de Excel u otras aplicaciones.

©Alfaomega-RC Libros

Puede instalar una instancia de Analysis Services en uno de estos tres modos de servidor:

- Como instancia tabular, ejecutando modelos tabulares.

- Como una instancia multidimensional y de minería de datos, ejecutando cubos OLAP y modelos de minería de datos (es el valor predeterminado).

- Como PowerPivot para SharePoint, ejecutando modelos de datos PowerPivot y de Excel en SharePoint (PowerPivot para SharePoint es un motor de datos de nivel intermedio que carga, consulta y actualiza modelos de datos hospedados en SharePoint).

El mismo motor de datos tiene tres formas de usarlo. Tenga en cuenta que los modos de servidor se establecen durante la instalación y no se pueden cambiar posteriormente. Debe instalar una nueva instancia si necesita otro modo diferente.

Analysis Services proporciona tres métodos para crear un modelo semántico de Business Intelligence: tabular, multidimensional y PowerPivot. Las soluciones tabulares usan construcciones de modelado relacional como tablas y relaciones para modelar los datos y el motor de análisis en memoria para almacenarlos y calcularlos. Las soluciones multidimensionales y de minería de datos usan construcciones de modelado OLAP (cubos y dimensiones) y almacenamiento MOLAP, ROLAP u HOLAP. PowerPivot es una solución BI de autoservicio que permite a los analistas de negocios generar un modelo de datos analíticos en un libro de Excel mediante el complemento PowerPivot para Excel. PowerPivot usa también el motor de análisis en memoria en Excel y en SharePoint. Dado que usa las soluciones PowerPivot usan Excel tanto para el modelado de datos como para la representación, para implementar un libro en un servidor para el acceso a datos centralizados y controlados, se requiere SharePoint y Excel Services.

Las soluciones tabulares y multidimensionales se generan con SQL Server Data Tools y se han diseñado para proyectos BI corporativos que se ejecutan en una instancia independiente de Analysis Services. Ambas soluciones producen bases de datos analíticas de alto rendimiento que se integran con facilidad con informes de Reporting Services, Excel y otras aplicaciones BI desde aplicaciones de Microsoft y de otros fabricantes. Con todo, cada solución difiere en cómo se crea, se usa y se implementa. En este tema se exploran las diferencias, lo que le permite comparar e identificar la solución que mejor cumpla los requisitos del proyecto.

Dado que la tabular es la solución más reciente, puede que piense que migrar una solución MDX existente a un formato tabular lo más correcto, pero este no suele ser el caso. La solución tabular no reemplaza a la multidimensional y los dos formatos no son intercambiables. A menos que tenga una razón concreta para ello, no recompile una solución MDX existente si cumple ya las necesidades de su organización. Para los proyectos nuevos, considere el método tabular. Se agiliza el diseño, la prueba y la implementación; y funcionará mejor con las aplicaciones BI de autoservicio más recientes de Microsoft.

Los **modelos tabulares son bases de datos "en memoria"** de Analysis Services. Gracias a los algoritmos de compresión avanzados y al procesador de consultas multiproceso, el motor analítico en memoria xVelocity (VertiPaq) ofrece un acceso rápido a los objetos y los datos de los modelos tabulares para aplicaciones cliente de informes como Microsoft Excel y Microsoft Power View.

Los modelos tabulares admiten el acceso a los datos mediante dos modos: modo de almacenamiento en caché y modo DirectQuery. En el modo de almacenamiento en caché, puede integrar datos de varios orígenes como bases de datos relacionales, fuentes de distribución de datos y archivos de texto planos. En el modo DirectQuery, puede omitir el modelo en memoria, lo que permite a las aplicaciones cliente consultar los datos directamente en el origen relacional (SQL Server).

Los modelos tabulares se crean en SQL Server Data Tools (SSDT) mediante las nuevas plantillas de proyectos de modelos tabulares. Puede importar datos de varios orígenes y, a continuación, enriquecer el modelo agregando relaciones, columnas calculadas, medidas, KPI y jerarquías. A continuación, los modelos se pueden implementar en una instancia de Analysis Services que permite a las aplicaciones cliente de informes conectarse con ellos. Los modelos implementados se pueden administrar en SQL Server Management Studio del mismo modo que los modelos multidimensionales. También se pueden crear particiones de los mismos para optimizar el procesamiento y protegerlos en el nivel de fila usando la seguridad basada en roles.

Una **solución multidimensional** de Analysis Services usa estructuras de cubos para analizar datos de negocio en varias dimensiones. El modo multidimensional es el modo de servidor predeterminado de Analysis Services. Incluye un motor de cálculo y consulta de datos OLAP, con los modos de almacenamiento MOLAP, ROLAP y HOLAP para equilibrar el rendimiento con los requisitos de datos escalables. El motor OLAP

 ©Alfaomega-RC Libros

de Analysis Services es un servidor OLAP lider en el sector, que funciona con una amplia variedad de herramientas de BI. La mayoría de las implementaciones de Analysis Services se instalan como servidores OLAP clásicos.

La razón principal para generar un modelo multidimensional de Analysis Services es lograr el rendimiento rápido de consultas ad hoc en los datos empresariales. Un modelo multidimensional se compone de cubos y dimensiones que se pueden anotar y ampliar para admitir construcciones de consultas complejas. Los desarrolladores de BI crean cubos para admitir tiempos de respuesta rápida y para proporcionar un único origen de datos para informes empresariales. Debido a la mayor importancia de business intelligence en todos los niveles de una organización, el hecho de tener un solo origen de datos analíticos garantiza que las discrepancias se mantienen al mínimo, si no se eliminan por completo.

Otra ventaja importante del uso de las bases de datos multidimensionales de Analysis Services es la integración con las herramientas de informes BI utilizadas habitualmente, como Excel, Reporting Services y PerformancePoint, así como las aplicaciones personalizadas y las soluciones de terceros.

Analysis Services también proporciona una plataforma integrada para las soluciones que incorporan la **minería de datos**. Puede usar datos relacionales o de cubo para crear soluciones de Business Intelligence con análisis predictivos.

La minería de datos usa principios estadísticos contrastados para detectar patrones en los datos, ayudándole a tomar decisiones inteligentes sobre problemas complejos. La aplicación de los algoritmos de minería de datos de Analysis Services a los datos le permitirá predecir tendencias, identificar patrones, crear reglas y recomendaciones, analizar la secuencia de eventos en conjuntos de datos complejos y obtener nuevos puntos de vista.

En SQL Server 2014, la minería de datos es eficaz y accesible, y está integrada con las herramientas preferidas de los usuarios para el análisis y la creación de informes. Vea los vínculos de esta sección para obtener toda la información sobre la minería de datos que necesita para empezar.

SQL Server proporciona las siguientes características para las soluciones integradas de minería de datos:

- **Varios orígenes de datos**: no es necesario crear un almacenamiento de datos o un cubo OLAP para realizar la minería de datos. Puede usar datos tabulares de proveedores externos, hojas de cálculo e incluso archivos de texto. También puede realizar fácilmente la minería de los cubos OLAP creados en

Analysis Services. Sin embargo, no puede usar datos de una base de datos en memoria.

- **Limpieza de los datos integrados, administración de datos y ETL**: Data Quality Services proporcionan herramientas avanzadas para la generación de perfiles y la limpieza de datos. Se puede usar Integration Services para generar procesos ETL de limpieza de datos, y también para tareas de creación, procesamiento, entrenamiento y actualización de modelos.

- **Varios algoritmos personalizables**: además de proporcionar algoritmos como la agrupación en clústeres, las redes neuronales y los árboles de decisión, la plataforma le permite desarrollar sus propios complementos con algoritmos personalizados.

- **Infraestructura de prueba del modelo**: pruebe los modelos y los conjuntos de datos usando herramientas estadísticas tan importantes como la validación cruzada, las matrices de clasificación, los gráficos de mejora respecto al modelo predictivo y los gráficos de dispersión. Cree y administre fácilmente conjuntos de prueba y entrenamiento.

- **Consultas y obtención de detalles**: cree consultas de predicción, recupere patrones y estadísticas de modelos, y obtenga información detallada de los datos de los casos.

- **Herramientas de cliente**: además de los estudios de desarrollo y diseño proporcionados por SQL Server, puede usar los Complementos de minería de datos para Excel para crear, consultar y examinar los modelos. O bien crear clientes personalizados, incluidos servicios web.

- **Compatibilidad con el lenguaje de scripting y API administrada**: todos los objetos de minería de datos son completamente programables. El scripting es posible mediante MDX, XMLA o las extensiones de PowerShell para Analysis Services. Use el lenguaje DMX (Extensiones de minería de datos) para crear rápidamente consultas y scripts.

- **Seguridad e implementación**: proporciona seguridad basada en roles a través de Analysis Services, incluyendo permisos distintos para la obtención de detalles del modelo y los datos de la estructura. Fácil implementación de modelos en otros servidores, de forma que los usuarios puedan tener acceso a los patrones o realizar predicciones.

©Alfaomega-RC Libros

CAPÍTULO 5

MODELADO PREDICTIVO CON HERRAMIENTAS DE IBM

TÉCNICAS DE MODELADO PREDICTIVO

La clasificación de las técnicas de análisis de datos discrimina entre la existencia o no de variables explicativas y explicadas.

Si existe una dependencia entre las variables explicadas y sus correspondientes variables explicativas, que pueda plasmarse en un ***modelo***, estamos ante las *técnicas predictivas* o *métodos explicativos o técnicas de modelado predictivo,* herramientas fundamentales en Business Intelligence.

Este tipo de técnicas de análisis de la dependencia pueden clasificarse en función de la naturaleza métrica o no métrica de las variables independientes y dependientes tal y como se indica en la Figura 5-1.

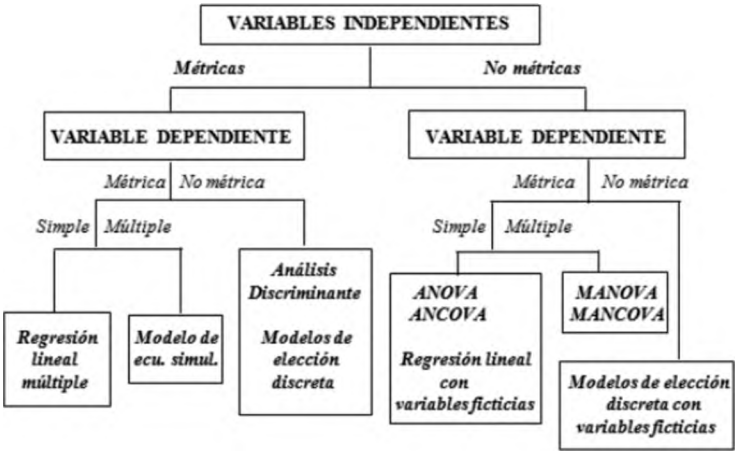


Figura 5-1

El ***análisis de la regresión múltiple*** es una técnica estadística utilizada para analizar la relación entre una variable dependiente (o endógena) métrica y varias variables independientes (o exógenas) también métricas. El objetivo esencial del análisis de la regresión múltiple es utilizar las variables independientes, cuyos valores son conocidos, para predecir la única variable criterio (dependiente) seleccionada por el investigador.

La expresión funcional del análisis de la regresión múltiple es la siguiente:

$$y = F(x_1, x_2, \cdots, x_n)$$

donde inicialmente, tanto la variable dependiente y como las independientes x_i son métricas. Asimismo la regresión múltiple admite la posibilidad de trabajar con variables independientes no métricas si se emplean variables ficticias (***modelos de regresión con variables ficticias***) para su transformación en métricas.

Los ***modelos de ecuaciones simultáneas*** constituyen una técnica estadística utilizada para analizar la relación entre múltiples variables dependientes (o endógenas) métricas y varias variables independientes (o exógenas) también métricas. El objetivo esencial es utilizar las variables independientes, cuyos valores son conocidos, para predecir las variables criterio (dependientes) seleccionadas por el investigador.

©Alfaomega-RC Libros

La expresión funcional es la siguiente:

$$\vec{G}(y_1, y_2, \cdots, y_n) = \vec{F}(x_1, x_2, \cdots, x_n)$$

donde inicialmente, tanto las variables dependientes y_i como las independientes x_i son métricas. Se observa que este modelo es una ampliación del modelo de regresión múltiple al caso de varias variables dependientes.

El ***análisis discriminante*** es una técnica estadística utilizada para analizar la relación entre una variable dependiente (o endógena) no métrica (categórica) y varias variables independientes (o exógenas) métricas. El objetivo esencial del análisis discriminante es utilizar los valores conocidos de las variables independientes para predecir con qué categoría de la variable dependiente se corresponden. Así podremos predecir en qué categoría de riesgo crediticio se encuentra una persona, el éxito de un producto en el mercado, etc.

La expresión funcional del análisis discriminante es la siguiente:

$$y = F(x_1, x_2, \cdots, x_n)$$

donde la variable dependiente y es no métrica y las variables independientes son métricas. Se trata por tanto de un caso particular del análisis de regresión múltiple. Formalmente podríamos decir que el análisis discriminante es una técnica de clasificación que permite agrupar a los elementos de una muestra en dos o más categorías diferentes, predefinidas en una variable dependiente no métrica, en función de una serie de variables independientes métricas combinadas linealmente.

En el análisis discriminante, para valores dados de las variables independientes hemos de predecir la probabilidad de pertenencia a una categoría o clase de la variable dependiente (por ejemplo, probabilidad de que un individuo compre un producto o devuelva un crédito según algunas variables medidas en él). Los ***modelos de elección discreta*** tienen la misma naturaleza que el modelo discriminante, pero ahora lo que se predice es la probabilidad de pertenencia a una categoría (clase) para valores dados de las variables dependientes. Por tanto, los modelos de elección discreta predicen directamente la probabilidad de ocurrencia de un suceso que viene definido por los valores de las variables independientes. Como los valores de una probabilidad están entre cero y uno, las predicciones realizadas con los modelos de elección discreta deben estar acotadas para que caigan en el rango entre cero y uno. El modelo general que

cumple esta condición es un caso particular del modelo de regresión múltiple que se denomina ***modelo lineal de probabilidad***, y tiene la forma funcional:

$$P_i = F(x_i, \beta) + u_i$$

Se observa que si *F* es la función de distribución de una variable aleatoria, entonces *P* varía entre cero y uno.

En el caso particular en que la función *F* es la función logística estaremos ante el ***modelo Logit o Regresión Logística***, cuya forma funcional será la siguiente:

$$P_i = F(x_i, \beta) + u_i = \frac{e^{x_i \beta}}{1 + e^{x_i \beta}} + u_i$$

En el caso particular en que la función *F* es la función de distribución de una normal unitaria estaremos ante el ***modelo Probit***, cuya forma funcional será la siguiente:

$$P_i = F(x_i, \beta) + u_i = (2\pi)^{-\frac{1}{2}} \int_{-\infty}^{x_i \beta} e^{-\frac{t^2}{2}} dt + u_i$$

El ***análisis de la varianza simple*** es una técnica estadística utilizada para analizar la relación entre una variable dependiente (o endógena) métrica y varias variables independientes (o exógenas) no métricas. El objetivo esencial de los modelos del análisis de la varianza es determinar si diversas muestras proceden de poblaciones con igual media. Los valores no métricos de las variables independientes determinarán una serie de grupos en la variable dependiente. De modo que el modelo ANOVA mide la significación estadística de las diferencias entre las medias de los grupos determinados en la variable dependiente por los valores de las variables independientes.

La expresión funcional del modelo del análisis de la varianza simple ANOVA es la siguiente:

$$y = F(x_1, x_2, \cdots, x_n)$$

donde la variable dependiente *y* es métrica y las variables independientes son no métricas. Se trata por tanto de otro caso particular del modelo de regresión múltiple.

El ***análisis de la covarianza simple*** es una técnica estadística utilizada para analizar la relación entre una variable dependiente (o endógena) métrica y varias variables independientes (o exógenas), parte de las cuales son no métricas, siendo la otra parte métricas (*covariables*).

 ©Alfaomega-RC Libros

La expresión funcional del modelo del análisis de la covarianza simple ANCOVA es la siguiente:

$$y = F(x_1, x_2, \cdots, x_n)$$

donde la variable dependiente *y* es métrica y las variables independientes son algunas métricas y otras no métricas. Se trata por tanto de otro caso particular del modelo de regresión múltiple.

El ***análisis de la varianza múltiple*** es una técnica estadística utilizada para analizar la relación entre varias variables dependientes (o endógenas) métricas y varias variables independientes (o exógenas) no métricas. El objetivo esencial de los modelos del análisis de la varianza múltiple es contrastar si los valores no métricos de las variables independientes determinarán la igualdad de vectores de medias de una serie de grupos determinados por ellos en las variables dependientes. De modo que el modelo MANOVA mide la significación estadística de las diferencias entre los vectores de medias de los grupos determinados en las variables dependientes por los valores de las variables independientes.

La expresión funcional del modelo del análisis de la varianza múltiple MANOVA es la siguiente:

$$G(y_1, y_2, \cdots, y_m) = F(x_1, x_2, \cdots, x_n)$$

donde las variables dependientes son métricas y las variables independientes son no métricas. Seguimos hablando de una caso particular de la regresión múltiple.

El ***análisis de la covarianza múltiple*** es una técnica estadística utilizada para analizar la relación entre varias variables dependientes (o endógenas) métricas y varias variables independientes (o exógenas) mezcla de variables métricas y no métricas.

La expresión funcional del modelo del análisis de la covarianza múltiple MANCOVA es la siguiente:

$$G(y_1, y_2, \cdots, y_m) = F(x_1, x_2, \cdots, x_n)$$

donde las variables dependientes son métricas y las variables independientes son una parte métricas y otra parte no métricas.

En el análisis de la covarianza, tanto simple como múltiple, las variables métricas independientes (*covariables*) tienen como objetivo eliminar determinados efectos que puedan sesgar los resultados incrementando la varianza dentro de los grupos. En el análisis de la covarianza se suele comenzar eliminando, mediante una regresión lineal, la variación experimentada por las variables dependientes producida por la covariable o covariables de efectos indeseados, para continuar con un análisis ANOVA o MANOVA sobre las variables dependientes ajustadas (residuos de la regresión anterior).

La regresión múltiple admite la posibilidad de trabajar con variables independientes no métricas si se emplean variables ficticias para su transformación en métricas. A cada clase de la variable no métrica se le asigna un valor numérico.

El ***modelo de regresión múltiple con variables ficticias*** es similar al análisis de la regresión múltiple con la diferencia de que las variables independientes pueden ser también no métricas. Por tanto, se trata de una técnica estadística utilizada para analizar la relación entre una variable dependiente (o endógena) métrica y varias variables independientes (o exógenas) métricas, no métricas o mezcla de ambas. El objetivo esencial del análisis de la regresión múltiple es utilizar las variables independientes, cuyos valores son conocidos, para predecir la única variable criterio (dependiente) seleccionada por el investigador.

La expresión funcional del análisis de la regresión múltiple con variables ficticias es la siguiente:

$$y = F(x_1, x_2, \cdots, x_n)$$

Al igual que la regresión múltiple, los modelos de elección discreta admiten la posibilidad de trabajar con variables independientes no métricas si se emplean variables ficticias para su transformación en métricas.

Los ***árboles de decisión*** constituyen una técnica estadística utilizada para analizar la relación entre una variable dependiente (o endógena) no métrica (categórica) y varias variables independientes (o exógenas) también no métricas. El objetivo esencial del árbol de decisión es utilizar los valores conocidos de las variables independientes para predecir con qué categoría de la variable dependiente se corresponden. Así podremos predecir en qué categoría de riesgo crediticio se encuentra una persona, el éxito de un producto en el mercado, etc.

 ©Alfaomega-RC Libros

La expresión funcional del análisis discriminante es la siguiente:

$$y = F(x_1, x_2, \cdots, x_n)$$

donde la variable dependiente *y* es no métrica y las variables independientes son también no métricas. Se trata por tanto de un caso particular del análisis de regresión múltiple. En los árboles de decisión también suelen utilizarse variables independientes métricas agrupando sus valores en pocas clases (no más de 5).

Realmente es muy interesante observar que todas las técnicas predictivas para la modelización expuestas hasta aquí son un caso particular o una extensión del modelo de regresión múltiple.

Podríamos tabular los *métodos predictivos (técnicas de la dependencia), según la naturaleza de sus variables dependientes e independientes*, como se indica en la Figura 5-2.

TÉCNICA	*Variables dependientes*	*Variables independientes*
ANOVA y MANOVA	Métrica (métricas)	No métricas
ANCOVA y MANCOVA	Métrica (métricas)	Métricas y no métricas
REGRESIÓN MÚLTIPLE	Métrica	Métricas
REGRESIÓN MÚLTIPLE (VARIABLES FICTICIAS)	Métrica	Métricas y no métricas
ECUACIONES SIMULTÁNEAS	Métricas y no métricas	Métricas y no métricas
DISCRIMINANTE	No métrica	Métricas
ELECCIÓN DISCRETA (VARIABLES FICTICIAS)	No métrica	Métricas y no métricas
ÁRBOL DE DECISIÓN	No métrica	No métricas

Figura 5-2

TÉCNICAS DE MODELADO CON IBM SPSS MODELER

IBM SPSS MODELER agrupa todas las técnicas de modelización en la ficha *Modelado* en la categoría *Todas* (Figura 5-3). Adicionalmente, el programa clasifica las técnicas de modelado en cuatro categorías:

Automático: engloba las técnicas de modelización automática (Figura 5-4).
Clasificación: engloba las técnicas predictivas (Figura 5-5).
Asociación: engloba las técnicas de asociación (Figura 5-6).
Segmentación: engloba las técnicas de segmentación (Figura 5-7).

Figura 5-3

Figura 5-4

Figura 5-5

Figura 5-6

Figura 5-7

TÉCNICAS DE MODELADO PREDICTIVO CON IBM SPSS MODELER

Ya sabemos que IBMSPSS Modeler agrupa las técnicas de modelización predictivas en la ficha *Modelado* categoría *Clasificación* (Figura 5-8).

Figura 5-8

©Alfaomega-RC Libros

La categoría *Clasificación* de la ficha *Modelado* contiene los siguientes nodos predictivos importantes:

- *Lineal*: construye un modelo predictivo lineal automático.
- *Regresión lineal*: permite crear y estimar un modelo de regresión lineal simple o múltiple.
- *Red neuronal*: construye un modelo predictivo de red neuronal.
- *Crear C5.0*: permite construir árboles de decisión y conjunto de reglas utilizando el algoritmo C5.0.
- *Discriminante*: construye un modelo predictivo discriminante.
- *Logística:* construye un modelo predictivo de regresión logística.
- *GenLin*: construye un modelo predictivo de regresión lineal generalizado.
- *Cox*: construye un modelo predictivo de regresión de Cox.
- *Arbol C&R*: construye un modelo de árbol Cart.
- *Quest*: construye un modelo de árbol Quest.
- CHAID: construye un modelo de árbol Chaid.

El nodo Regresión lineal: Modelo de regresión múltiple

El nodo *Regresión Lineal* de la paleta *Modelado* permite crear y estimar un modelo de regresión lineal simple o múltiple. Como ejemplo se consideran 9 variables medidas sobre 100 madres y sus hijos recién nacidos en parto normal contenidas en el conjunto de datos ASCII de nombre *princip.txt*. Las variables son peso de la madre (*PESOM*), talla de la madre (*TALLAM*), semanas de gestación (*SEM*), presión arterial sistólica de la madre (*PASM*), presión arterial diastólica de la madre (*PADM*), peso del recién nacido (*PESOR*), talla del recién nacido (*TALLAR*), perímetro torácico del recién nacido (*PTR*) y perímetro craneal del recién nacido (*PCR*). El objetivo es intentar ajustar un modelo lineal de regresión múltiple que explique la talla del recién nacido en función de la talla de la madre, el peso de la madre y las semanas de gestación.

Comenzamos utilizando el nodo *Archivo variable* de la categoría *Orígenes* para situar como origen de datos el fichero *princip.txt*.

Es una tarea previa al uso de cualquier modelo, definir el tipo de las variables que lo componen. El nodo *Tipo* permite asignar campos como variables dependientes o independientes en un modelo y no utilizar los que se desee. Para ello añadimos el nodo *Tipo* de la categoría *Operaciones con campos* al diagrama. A continuación hacemos clic con el botón derecho del ratón sobre el nodo *Tipo* y elegimos *Edición* en el menú emergente resultante. Se obtiene la pantalla *Tipo* con las características de todas las variables del archivo de datos. Haciendo clic sobre la celda correspondiente a cada variable en la columna *Papel*, situamos el papel de la variable TALLAR como *Objetivo* (por ser la variable dependiente del modelo de regresión) y el papel del resto de las variables como *Entrada* (Figura 5-9). Al pulsar *Aceptar* se memorizan los tipos de los campos.

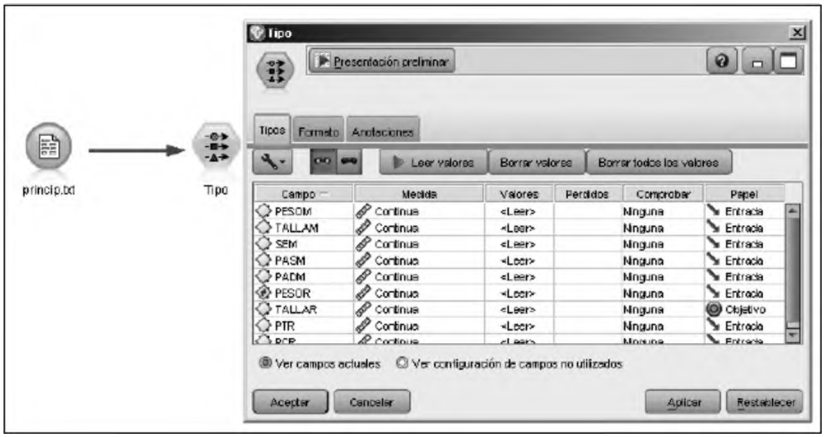


Figura 5-9

Para realizar el análisis de regresión múltiple añadimos el nodo *Regresión* de la categoría *Clasificación* de la ficha *Modelado* al diagrama que ya contiene el origen de datos *princip.txt*. A continuación hacemos clic con el botón derecho del ratón sobre el nodo *Regresión* y elegimos *Edición* en el menú emergente resultante. Se obtiene la pantalla de la Figura 5-10 en cuya solapa *Campos* definimos la variable dependiente del modelo (*Objetivo*) y las variables independientes (*Entradas*). En el campo *Método* de la solapa *Método* elegimos la forma de realizar la regresión (*Introducir, Por pasos, Hacia delante* o *Hacia atrás*). Utilizamos *Introducir* porque es el usuario quien selecciona las variables para la regresión (Figura 5-11). En el campo *Incluir constante en ecuación* se decide si el modelo tiene o no término constante. En la solapa *Experto* marcamos el campo *Experto* y hacemos clic en *Resultados* para marcar los resultados que deseamos en la regresión (Figura 5-12). Se observa que es posible elegir prácticamente la diagnosis completa del modelo de regresión lineal múltiple. Al pulsar *Aceptar* se memorizan las características del análisis de la regresión.

 ©Alfaomega-RC Libros

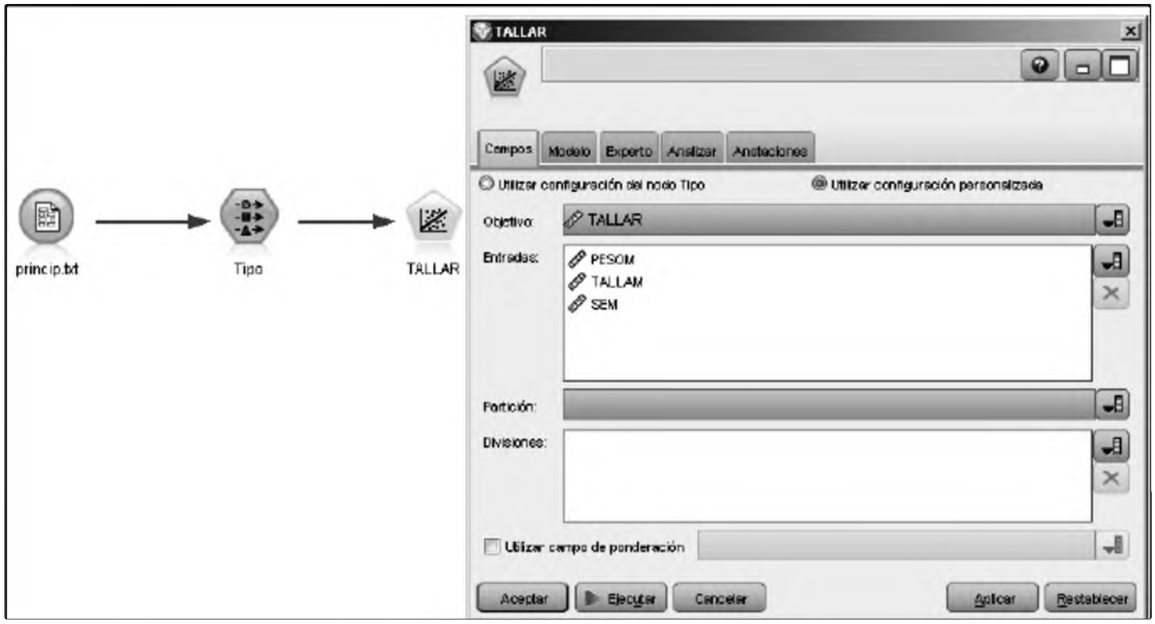


Figura 5-10

Figura 5-11

Figura 5-12

A continuación hacemos clic con el botón derecho del ratón sobre el nodo *Regresión* y elegimos *Ejecutar* en el menú emergente resultante o hacemos clic en el botón *Ejecutar* en la Figura 5-12. Se obtiene la pantalla de la Figura 5-13 que presenta un nuevo modelo generado de nombre *TALLAR* (variable independiente del modelo de regresión). Si hacemos clic con el botón derecho del ratón sobre el modelo *TALLAR* y elegimos *Edición* en el menú emergente resultante (Figura 5-14), se obtiene el modelo ajustado (Figura 5-15). La solapa *Modelo* nos da una idea de la importancia de cada variable independiente sobre la variable objetivo. Por ejemplo, al hacer clic sobre la barra relativa a la variable SEM se observa que su influencia sobre la variable objetivo para la predicción es de un 42%.

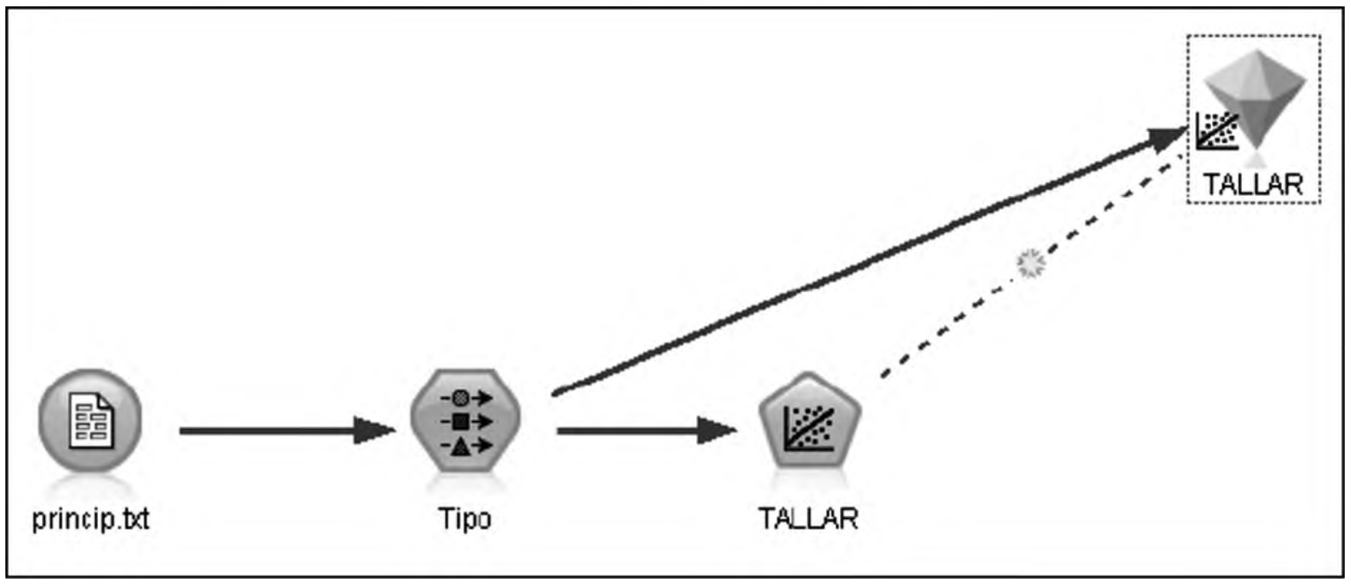


Figura 5-13

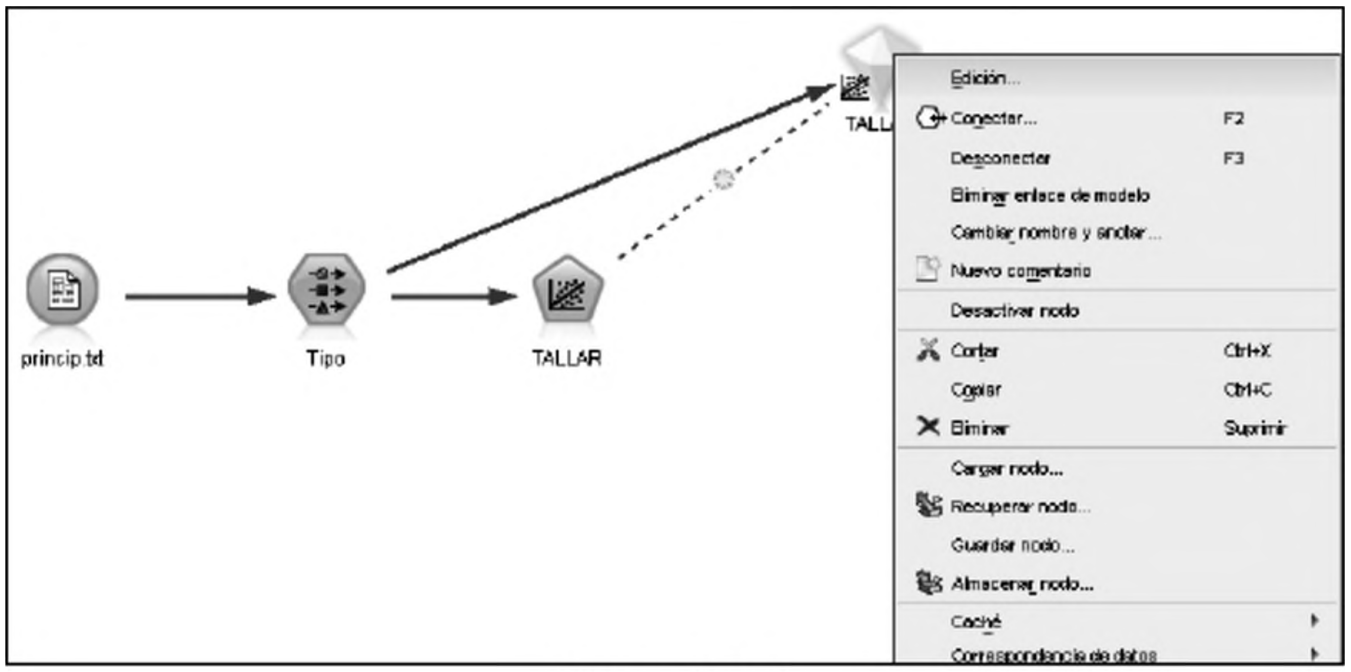


Figura 5-14

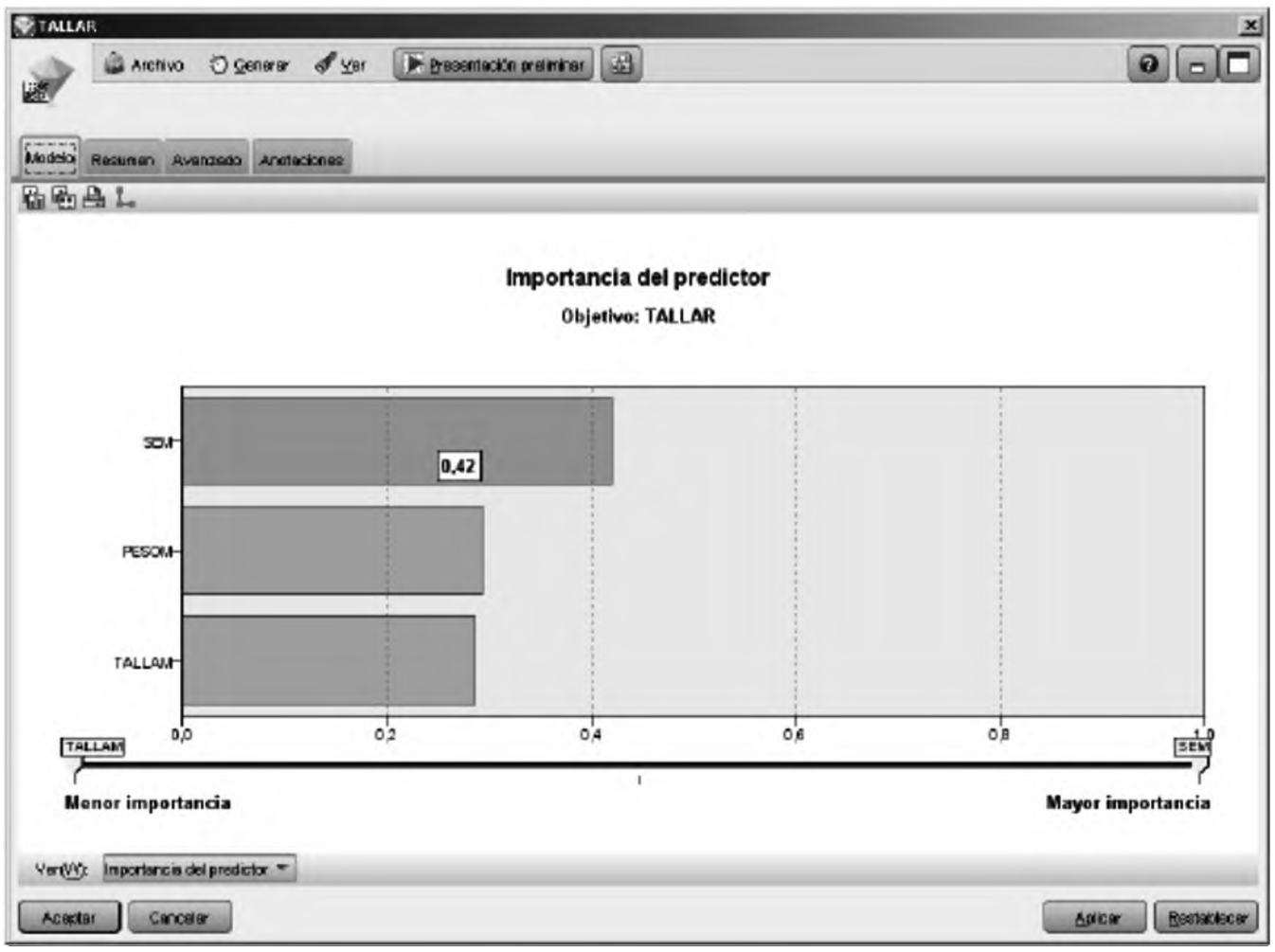


Figura 5-15

©Alfaomega-RC Libros

La solapa *Resumen* presenta la ecuación del modelo ajustado e información sobre las variables y el algoritmo utilizado en la regresión (Figura 5-16). La solapa *Avanzado* presenta los resultados de la regresión. En la Figura 5-17 se observan estadísticos y correlaciones de las variables del modelo.

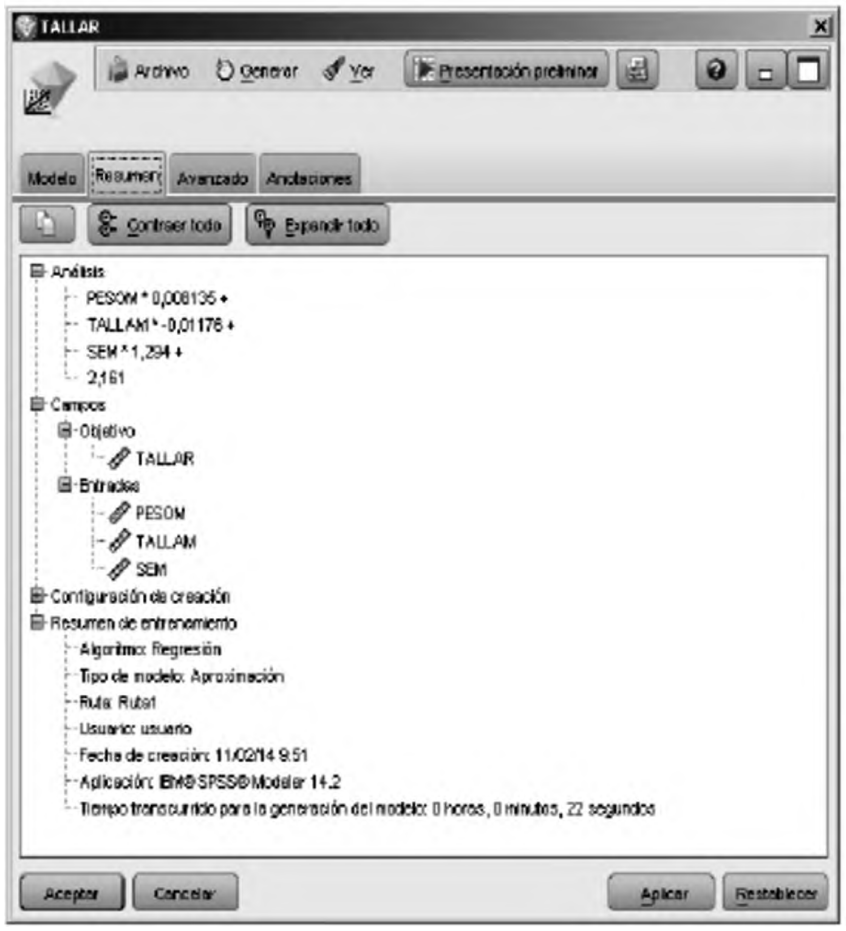


Figura 5-16

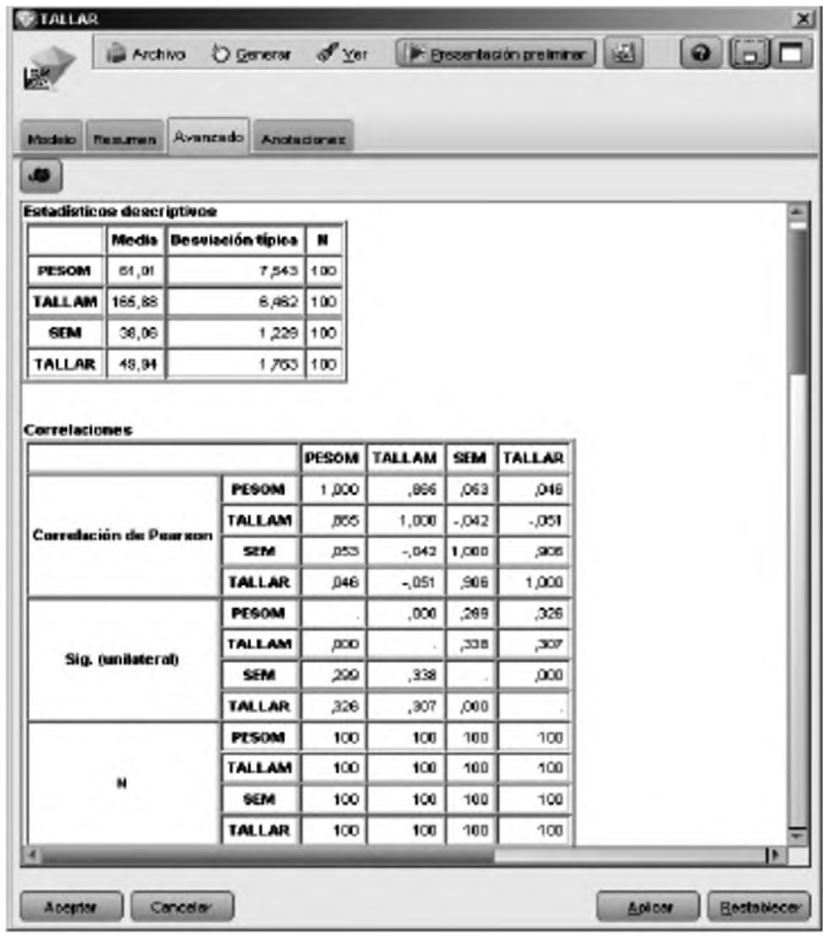


Figura 5-17

Variables introducidas/eliminadas(a)

Modelo	Variables introducidas	Variables eliminadas	Método
1	SEM, TALLAM, PESOM(b)		Introducir

a. Variable dependiente: TALLAR

b. Todas las variables solicitadas introducidas.

Resumen del modelo(a)

Modelo	R	R cuadrado	R cuadrado corregida	Error tip. de la estimación	Estadísticos de cambio					Criterio de selección				Durbin-Watson
					Cambio en R cuadrado	Cambio en F	gl1	gl2	Sig. Cambio en F	Criterio de información de Akaike	Criterio de predicción de Amemiya	Criterio de predicción de Mallows	Criterio bayesiano de Schwarz	
1	,906(b)	,821	,815	,757	,821	146,817	3	96	,000	-51,686	,194	4,000	-41,266	2,044

a. Variable dependiente: TALLAR

b. Variables predictoras: (Constante), SEM, TALLAM, PESOM

ANOVA(a)

Modelo		Suma de cuadrados	gl	Media cuadrática	F	Sig.
1	Regresión	252,587	3	84,196	146,817	,000(b)
	Residual	55,053	96	,573		
	Total	307,640	99			

a. Variable dependiente: TALLAR

b. Variables predictoras: (Constante), SEM, TALLAM, PESOM

Coeficientes(a)

Modelo		Coeficientes no estandarizados		Coeficientes tipificados	t	Sig.	Intervalo de confianza de 95,0% para B		Correlaciones			Estadísticos de colinealidad	
		B	Error tip.	Beta			Límite inferior	Límite superior	Orden cero	Parcial	Semiparcial	Tolerancia	FIV
1	(Constante)	2,161	4,078		,530	,597	-5,934	10,256					
	PESOM	,008	,020	,035	,398	,691	-,032	,049	,046	,041	,017	,244	4,096
	TALLAM	-,012	,024	-,043	-,494	,623	-,059	,036	-,051	-,050	-,021	,244	4,091
	SEM	1,294	,063	,902	20,539	,000	1,169	1,419	,906	,903	,887	,966	1,035

Figura 5-18

En la Figura 5-18 se observa un resumen del modelo que presenta el coeficiente de determinación R^2, el R^2 corregido (bastante bueno ya que el modelo explica el 81,5% de la variabilidad), el error típico de la estimación, el estadístico de Durbin Watson para el análisis de la autocorrelación (su valor cercano a 2 indica ausencia de autocorrelación en el modelo). También se observa la tabla ANOVA que muestra un p-valor de la *F* de Fisher Snedecor muy bajo (*Sig*. casi cero) indicando una significatividad conjunta de los parámetros estimados muy alta. En los resultados sobre coeficientes aparecen los parámetros estimados del modelo, su error típico, el valor de la *T* de Student y el *p*-valor asociado los intervalos de confianza, las correlaciones y el diagnóstico de colinealidad basados en la tolerancia y en el FIV (valores menores que 10 indican ausencia de problemas). Se observan p-valores muy altos para las variables *PESOM* y *TALLAM* y la constante, resultando significativa solamente la variable *SEM,* lo que indica que el resto de las variables deberían ser excluidas del modelo.

Diagnósticos de colinealidad(a)

Modelo	Dimensión	Autovalores	Índice de condición	Proporciones de la varianza			
				(Constante)	PESOM	TALLAM	SEM
1	1	3,989	1,000	,00	,00	,00	,00
	2	,010	19,941	,00	,25	,00	,02
	3	,001	71,730	,05	,14	,16	,68
	4	,000	148,323	,95	,61	,84	,30

a. Variable dependiente: TALLAR

Estadísticos sobre los residuos(a)

	Mínimo	Máximo	Media	Desviación típica	N
Valor pronosticado	45,96	53,75	49,94	1,597	100
Residual	-2,144	1,531	,000	,746	100
Valor pronosticado tip.	-2,493	2,384	,000	1,000	100
Residuo tip.	-2,831	2,022	,000	,985	100

a. Variable dependiente: TALLAR

Figura 5-19

En la Figura 5-19 aparecen los diagnósticos de colinealidad basados en los índices de condición y los estadísticos sobre los residuos. Como existen índices de condición superiores a 30 puede haber problemas de multicolinealidad. El modelo resulta tener bastantes problemas que lo invalidan.

El nodo Lineal: Modelo de regresión múltiple automático

El nodo *Lineal* ajusta automáticamente el modelo lineal de regresión múltiple más adecuado para las variables dadas. Como ejemplo ajustaremos el modelo anterior considerando como variable dependiente TALLAR y como posibles variables independientes el resto de las variables del fichero *princip.txt*. Comenzamos enlazando en la ruta el nodo *Lineal* de la categoría *Clasificación* de la ficha *Modelo* al nodo Tipo definido previamente en la Figura 5-9. Al hacer clic con el botón derecho del ratón sobre el nodo *Lineal* y elegir *Edición*, se obtiene la pantalla de la Figura 5-20. En la solapa *Campos* se observa la variable objetivo y las variables predictoras (entradas).

 ©Alfaomega-RC Libros

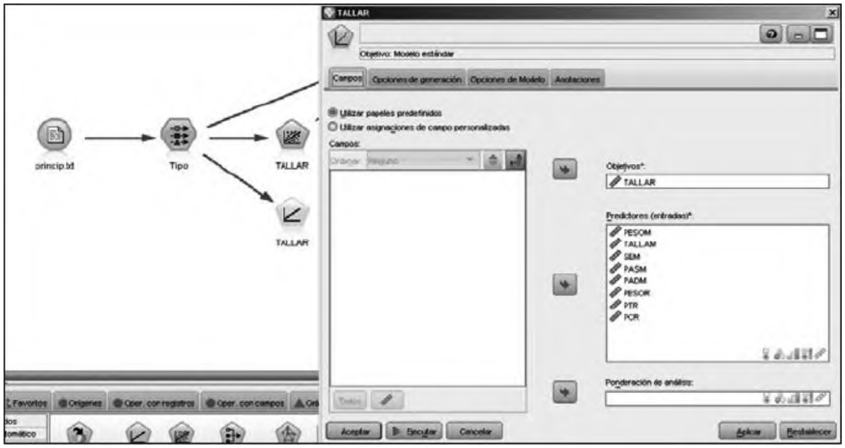


Figura 5-20

En la solapa *Opciones de predicción* (Figura 5-21) se observan una serie de resultados según las categorías de la izquierda de la pantalla. La categoría *Objetivos* permite elegir el modelo que se va a crear (estándar, boosting, bagging o conjunto de datos grande). La categoría *Procedimientos básicos* permite elegir que el programa prepare los datos automáticamente realizando el análisis exploratorio adecuado (Figura 5-22). La categoría permite elegir las características del modelo para el ajuste (Figura 5-23). La categoría *Conjuntos* permite gobernar la regresión para conjuntos de datos muy grandes (Figura 5-24). La categoría *Avanzados* permite hacer replicaciones aleatorias de resultados (Figura 5-25). La solapa *Opciones de Modelo* permite elegir el ajuste automático del modelo (Figura 5-26). Al hacer clic en *Ejecutar* se ejecuta la regresión automática y se obtiene el correspondiente modelo en la ruta (Figura 5-27).

Figura 5-21

Figura 5-22

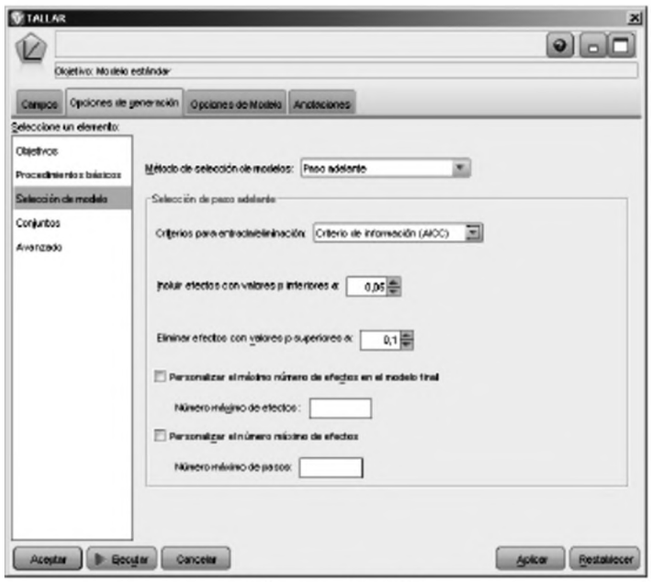


Figura 5-23

Figura 5-24

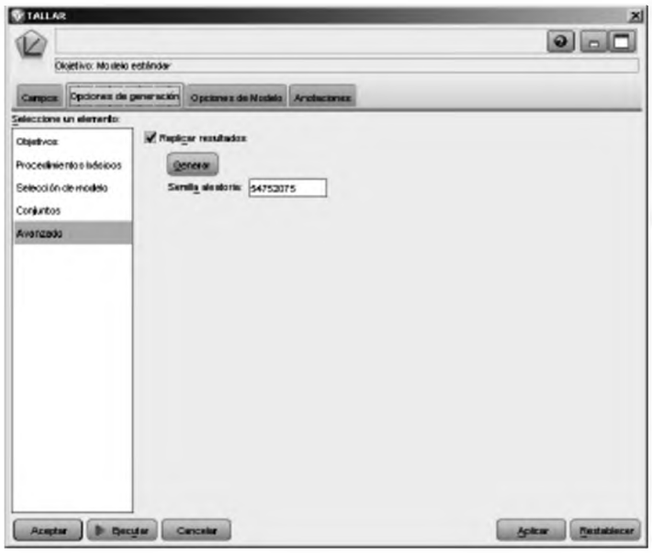


Figura 5-25

Figura 5-26

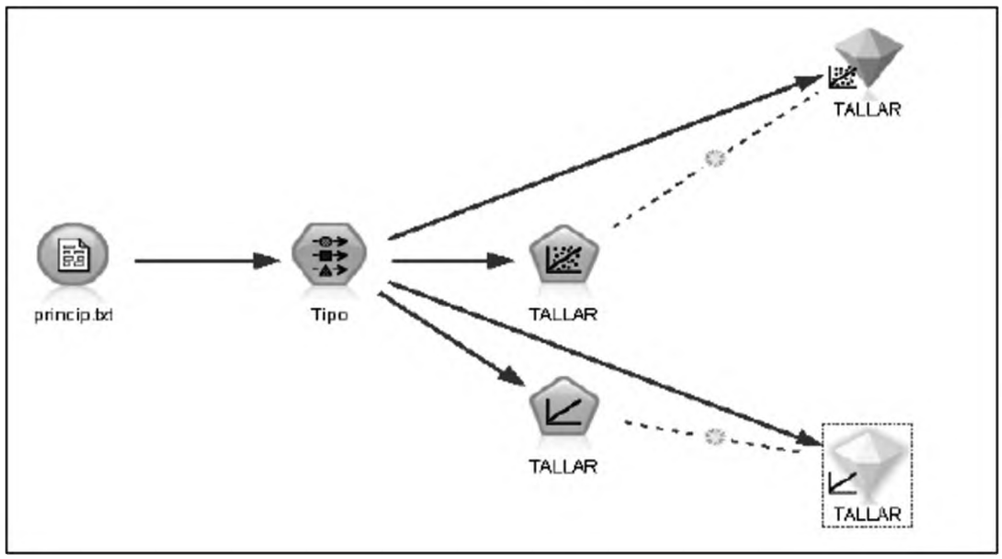


Figura 5-27

©Alfaomega-RC Libros

Si hacemos clic con el botón derecho del ratón sobre el modelo *TALLAR* y elegimos *Edición* en el menú emergente resultante, se obtiene el modelo ajustado (Figura 5-28). La solapa *Modelo* presenta una serie de paneles a la izquierda de la pantalla que van presentando los resultados de la regresión automática. El panel *Resumen del modelo* nos informa sobre el método de construcción del modelo y de las variables explicativas finalmente introducidas. El panel *Preparación de datos automática* nos informa de las acciones realizadas a las variables en el análisis exploratorio previo a la regresión (Figura 5-29). El panel *Importancia del predictor* nos informa de la influencia de cada variable explicativa sobre la variable dependiente (Figura 5-30). El panel *Pronóstico por observado* realiza un gráfico de valores predichos contra valores observados que debe ajustarse a la diagonal del primer cuadrante para que el ajuste sea bueno (Figura 5-31). El panel *Residuos* realiza un histograma de los residuos para analizar su normalidad (Figura 5-32). El panel *Valores atípicos* informa sobre los registros influyentes en la regresión por tener distancia de Cook alta (Figura 5-33). El panel *Efectos* presenta las variables independientes significativas en la regresión que en nuestro caso son SEM y PCR. Al situar el ratón sobre las líneas de las variables observamos el p-valor y un coeficeinte que da una idea de la importancia de cada variable independiente sobre la variable objetivo (Figuras 5-34 y 5-35). El panel *Coeficientes* presenta las variables independientes del modelo y la constante (si ha lugar) y al situar el ratón sobre la línea de cada variable vemos el valor del coeficiente estimado, el p-valor y la importancia de la variable (Figura 5-36). El panel *Resumen del modelo* (Figura 5-37) ofrece un resumen de los pasos dados para obtener la estimación del modelo, así como las variables significativas finalmente incorporadas al mismo.

Figura 5-28

Figura 5-29

Si hacemos clic con el botón derecho del ratón sobre el modelo *TALLAR* y elegimos *Edición* en el menú emergente resultante, se obtiene el modelo ajustado (Figura 5-28). La solapa *Modelo* presenta una serie de paneles a la izquierda de la pantalla que van presentando los resultados de la regresión automática. El panel *Resumen del modelo* nos informa sobre el método de construcción del modelo y de las variables explicativas finalmente introducidas. El panel *Preparación de datos automática* nos informa de las acciones realizadas a las variables en el análisis exploratorio previo a la regresión (Figura 5-29). El panel *Importancia del predictor* nos informa de la influencia de cada variable explicativa sobre la variable dependiente (Figura 5-30). El panel *Pronóstico por observado* realiza un gráfico de valores predichos contra valores observados que debe ajustarse a la diagonal del primer cuadrante para que el ajuste sea bueno (Figura 5-31). El panel *Residuos* realiza un histograma de los residuos para analizar su normalidad (Figura 5-32). El panel *Valores atípicos* informa sobre los registros influyentes en la regresión por tener distancia de Cook alta (Figura 5-33). El panel *Efectos* presenta las variables independientes significativas en la regresión que en nuestro caso son SEM y PCR. Al situar el ratón sobre las líneas de las variables observamos el p-valor y un coeficeinte que da una idea de la importancia de cada variable independiente sobre la variable objetivo (Figuras 5-34 y 5-35). El panel *Coeficientes* presenta las variables independientes del modelo y la constante (si ha lugar) y al situar el ratón sobre la línea de cada variable vemos el valor del coeficiente estimado, el p-valor y la importancia de la variable (Figura 5-36). El panel *Resumen del modelo* (Figura 5-37) ofrece un resumen de los pasos dados para obtener la estimación del modelo, así como las variables significativas finalmente incorporadas al mismo.

Figura 5-28

Figura 5-29

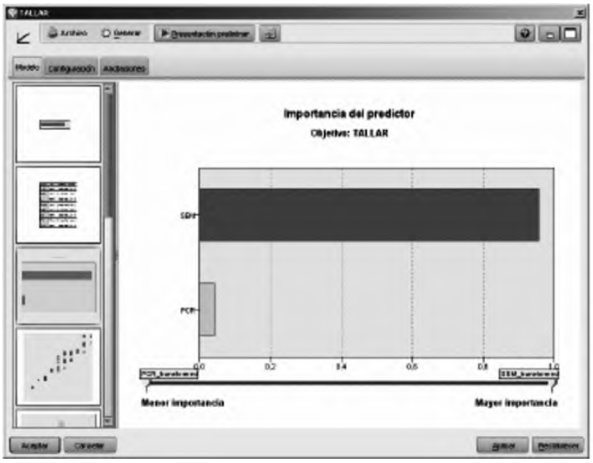


Figura 5-30

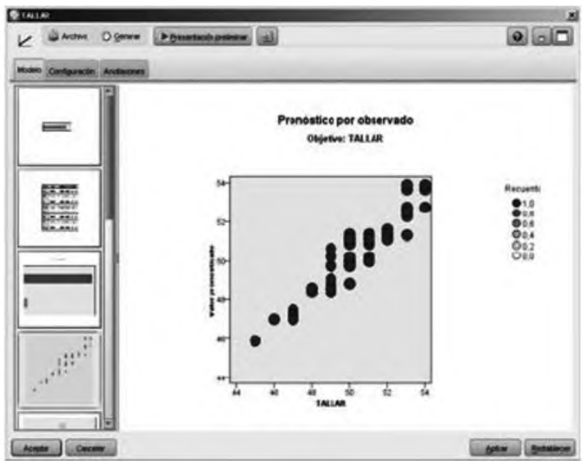


Figura 5-31

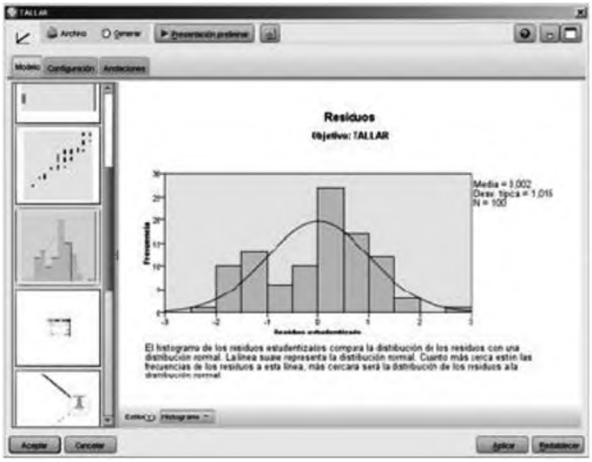


Figura 5-32

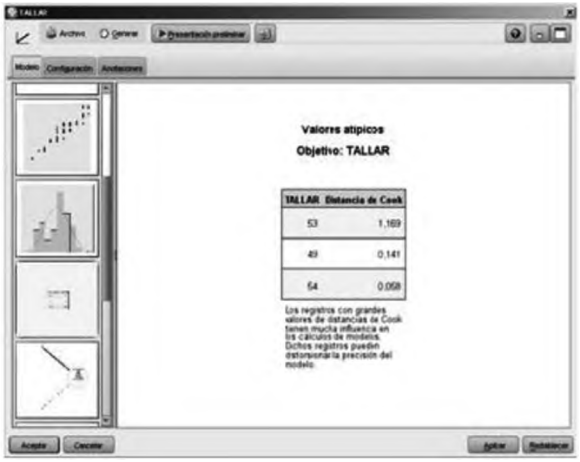


Figura 5-33

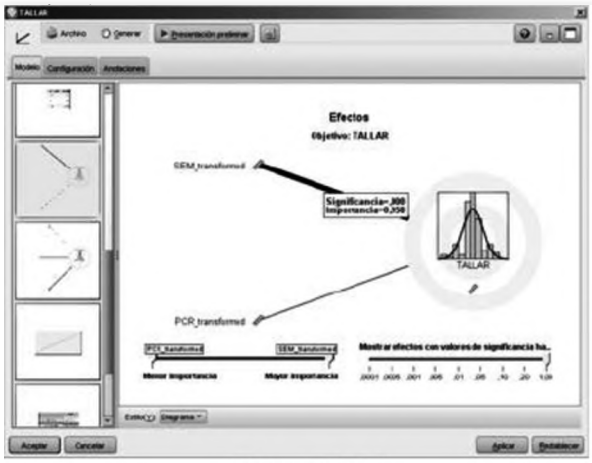


Figura 5-34

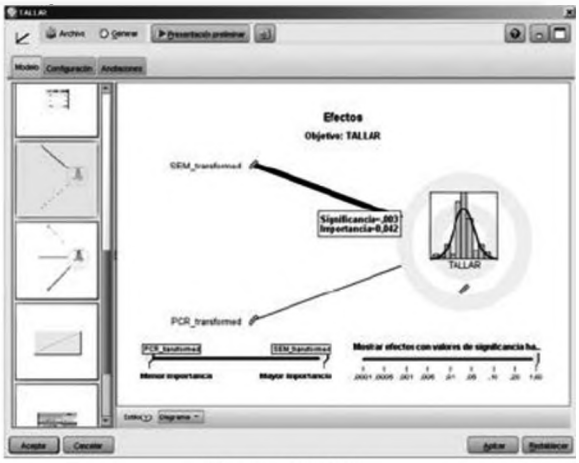


Figura 5-35

©Alfaomega-RC Libros

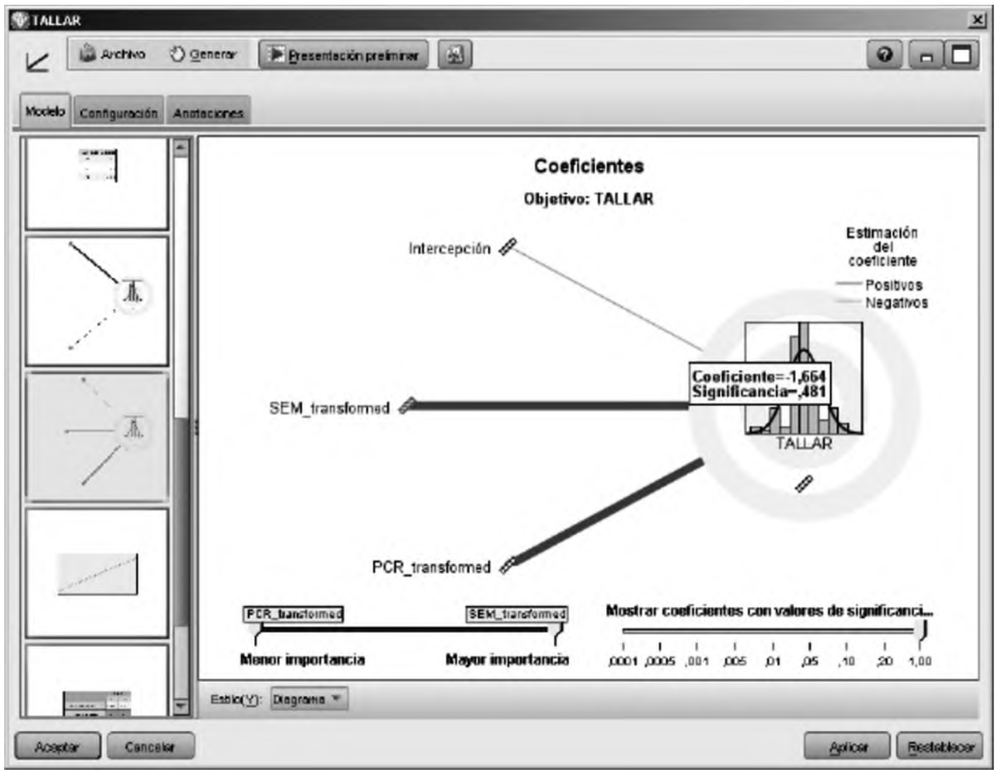


Figura 5-36

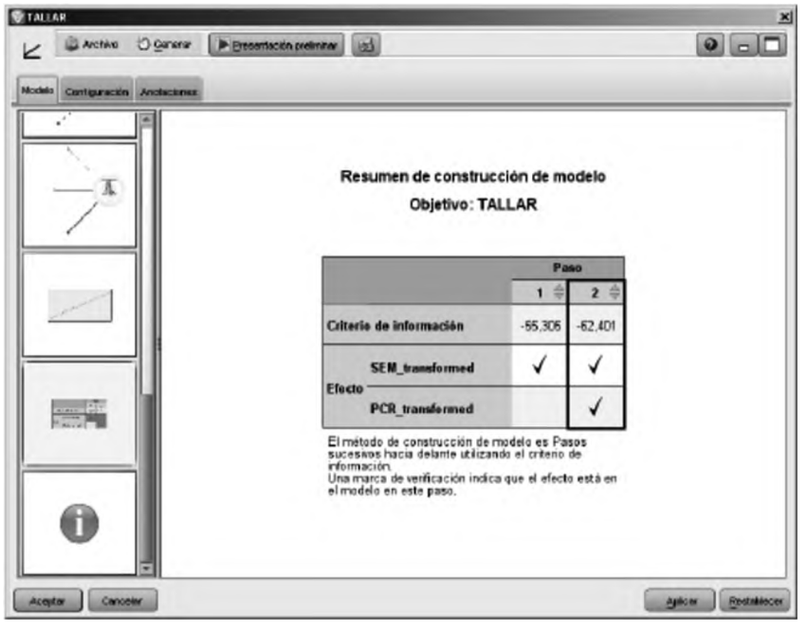


Figura 5-37

El nodo Regresión Logística: Modelos de elección discreta

El nodo *Regresión Logística* de la categoría *Clasificación* de la ficha *Modelado* permite crear y ajustar modelos de regresión logística con la finalidad de clasificar registros. Como ejemplo se considera el conjunto de datos ASCII de nombre *drug1n.txt*. Se trata de ajustar un modelo logístico que estime la probabilidad de tener colesterol en función del sexo y la droga aplicada.

Comenzamos utilizando el nodo *Archivo variable* de la categoría *Orígenes* para situar como origen de datos el fichero *drug1n.txt*. A continuación se añade el nodo *Tipo* de la categoría *Operaciones con campos* al diagrama y se sitúan como campos de entrada *Sexo* y *Droga*, y como campo objetivo *Colesterol* (Figura 5-38). Para realizar el análisis de regresión logística añadimos el nodo *Logística* de la categoría *Clasificación* de la ficha *Modelado* al diagrama que ya contiene el origen de datos *drug1n.txt*. A continuación hacemos clic con el botón derecho del ratón sobre el nodo *Logística* y elegimos *Edición* en el menú emergente resultante. Se obtiene la pantalla de la Figura 5-39. En la solapa *Campos* elegimos *Utilizar configuración personalizada* e introducimos la variable *Objetivo* y las variables de *Entrada* del modelo logístico. En la solapa *Modelo* elegimos *Automático*, el método *Introducir*, regresión logística *Multinomial*, modelo con solo *Efectos principales* e *Incluir constante en ecuación* (Figura 5-40). En la solapa *Experto* en el campo *Modo* elegimos *Experto* (Figura 5-41). En la solapa *Analizar* elegimos *Calcular importancia del predictor* en el campo *Evaluación del modelo* y *Calcular puntuaciones brutas de propensión* en el campo *Puntuaciones de propensión* (Figura 5-42). En el campo *Anotaciones* elegimos *Automático* (Figura 5-43) y señalamos aquellas opciones de salida que nos interesen. Al pulsar *Aceptar* se memorizan las características del análisis de la regresión logística y al hacer clic en *Ejecutar* en la parte inferior de la pantalla *Colesterol*, se realiza la regresión logística y aparece el modelo de la regresión logística en la ruta etiquetado con la variable dependiente (objetivo) del modelo (Figura 5-44).

Figura 5-38

©Alfaomega-RC Libros

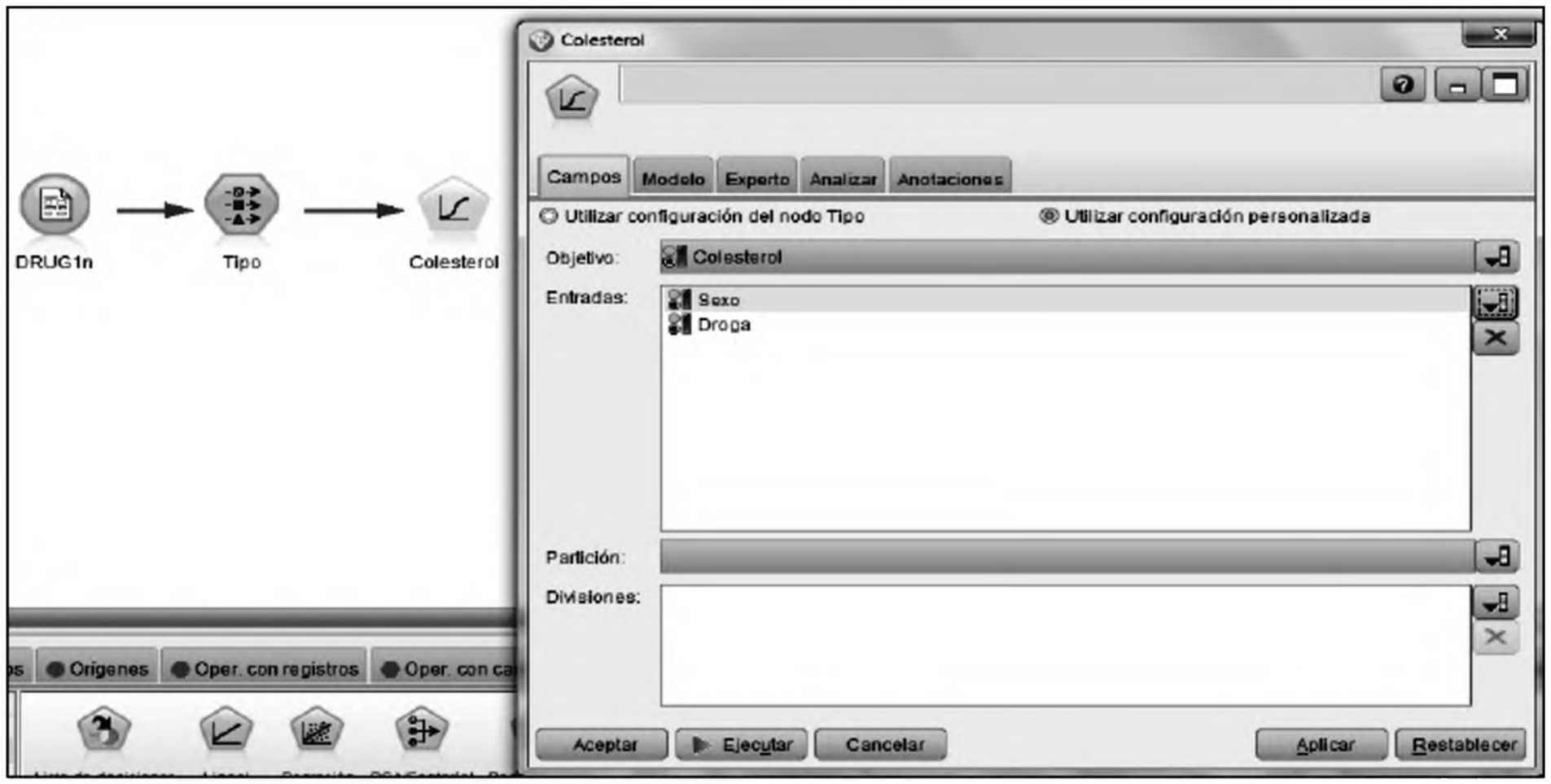


Figura 5-39

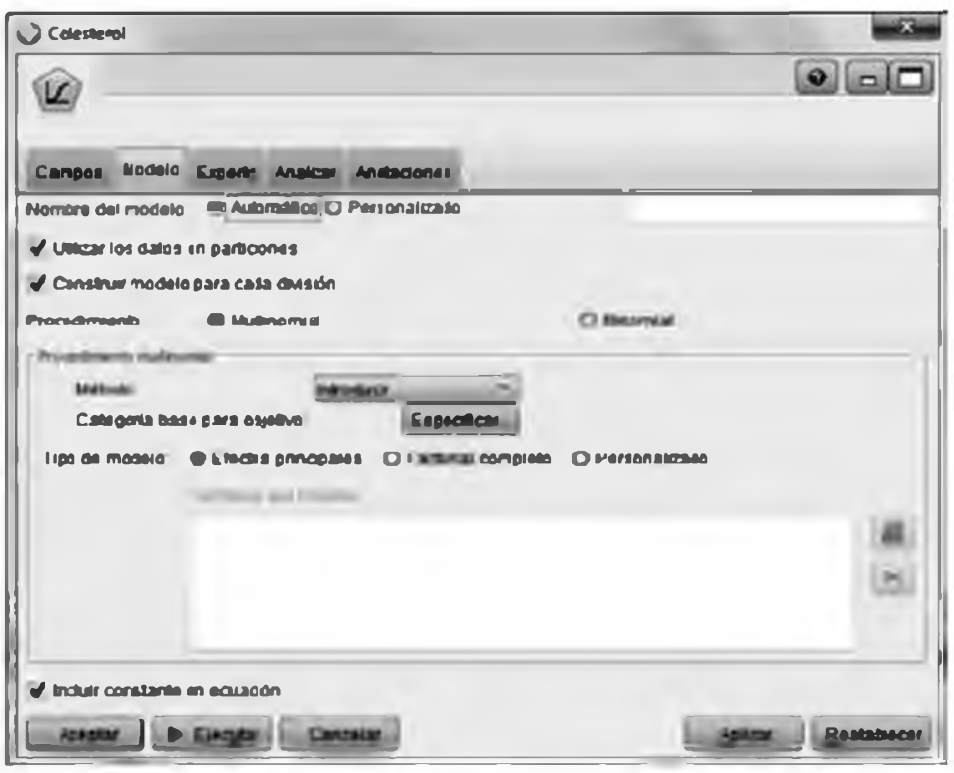


Figura 5-40

Figura 5-41

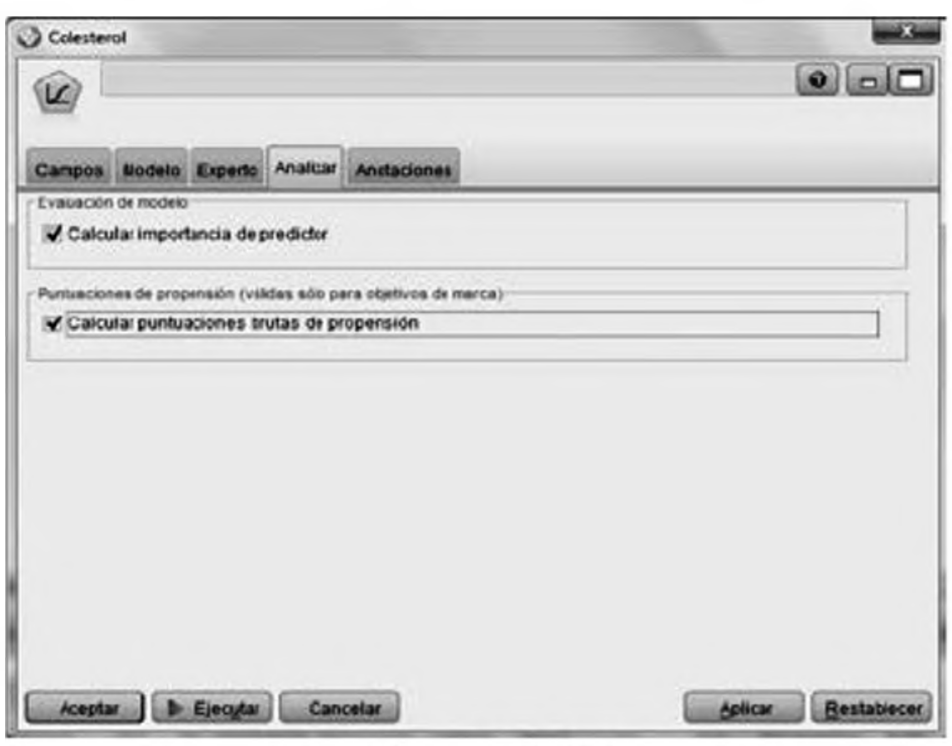


Figura 5-42

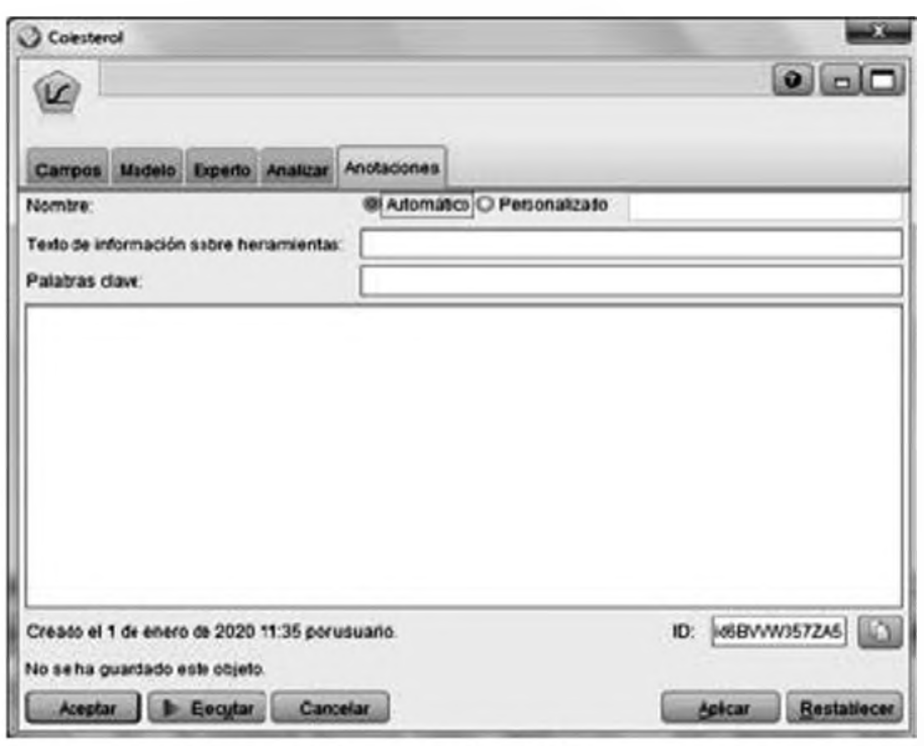


Figura 5-43

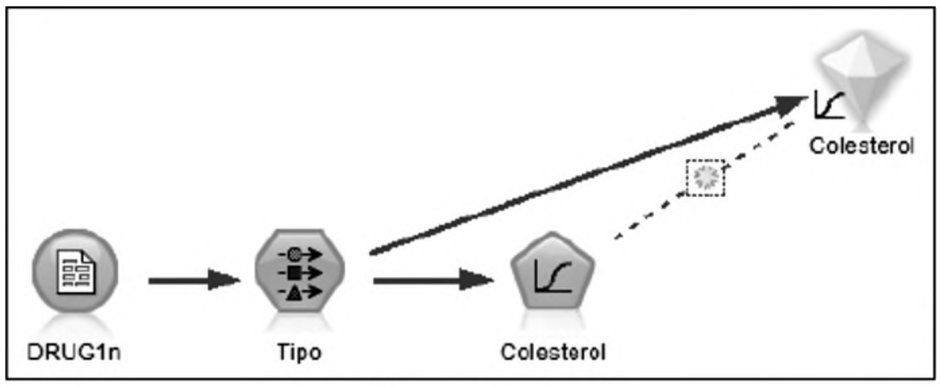


Figura 5-44

Si hacemos clic con el botón derecho del ratón sobre el modelo *Regresión* y elegimos *Edición* en el menú emergente resultante (Figura 5-45) obtenemos la pantalla de resultados de la regresión logística (Figura 5-46). En la solapa *Modelo* vemos la ecuación del modelo ajustado y la importancia de los predictores sobre la variable objetivo (Figura 5-46). En la solapa *Resumen* vemos un resumen del modelo (Figura 5-47). En la solapa *Avanzado* vemos el ajuste del modelo (Figura 5-48).

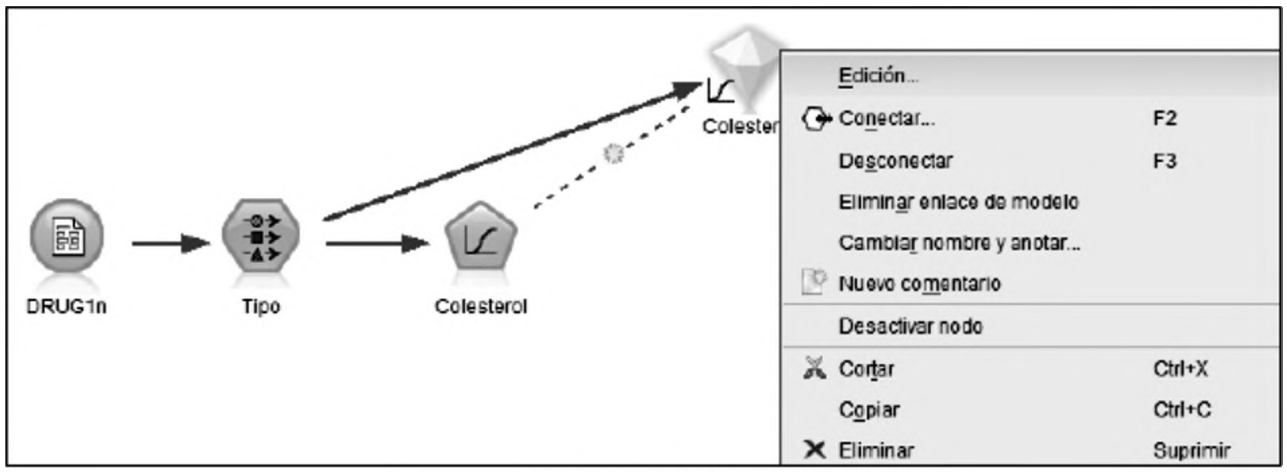


Figura 5-45

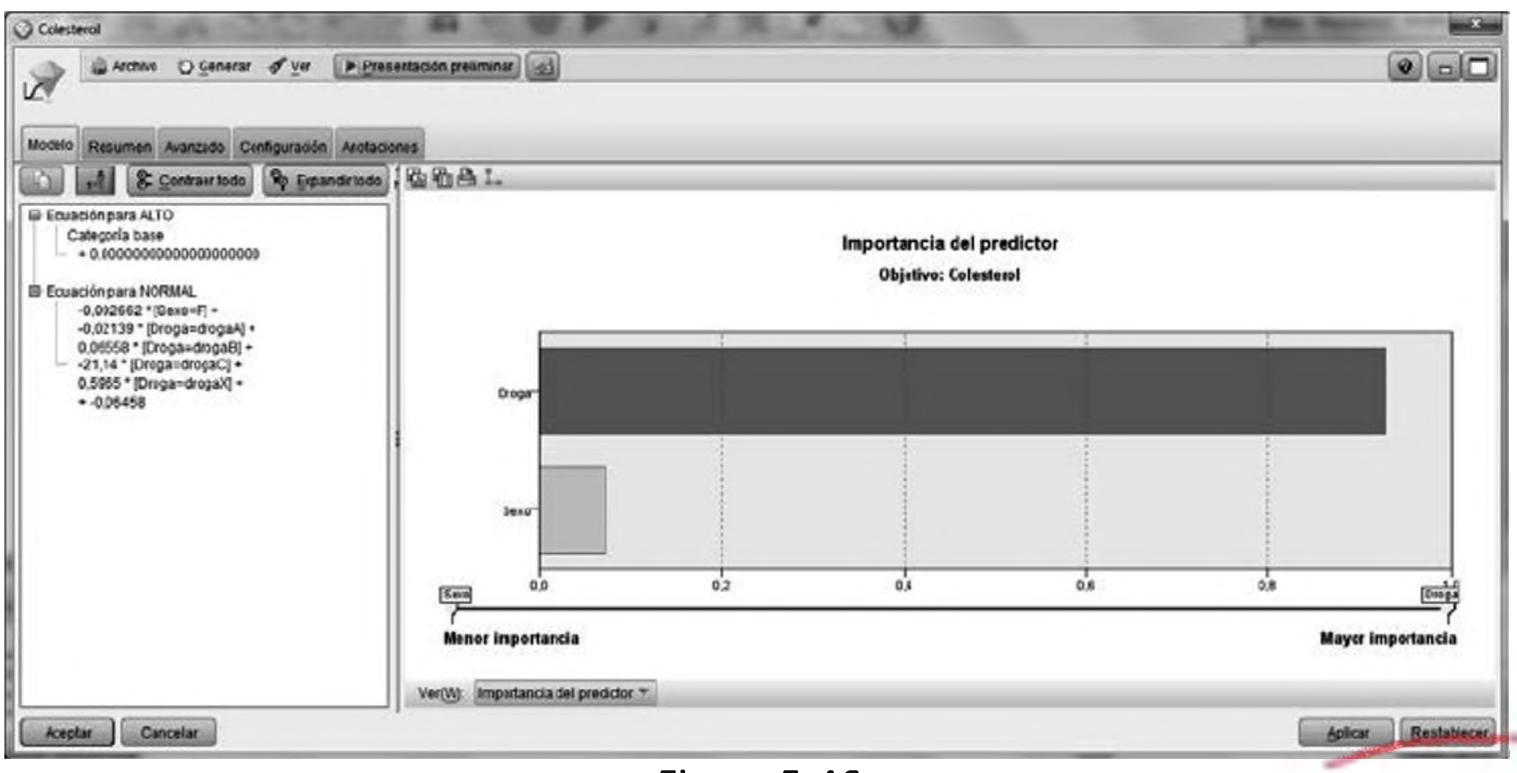


Figura 5-46

©Alfaomega-RC Libros

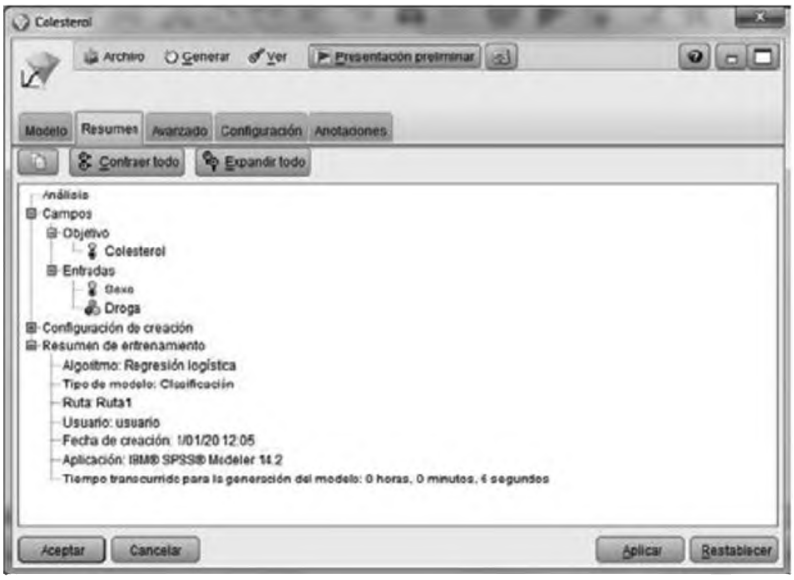


Figura 5-47

Información del ajuste del modelo

Modelo	Criterio de ajuste del modelo			Contrastes de la razón de verosimilitud		
	AIC	BIC	-2 log verosimilitud	Chi-cuadrado	gl	Sig.
Sólo la intersección	58,921	62,219	56,921			
Final	43,106	62,896	31,106	25,815	5	,000

Pseudo R-cuadrado

Cox y Snell	,121
Nagelkerke	,162
McFadden	,093

Estimaciones de los parámetros

Colesterol(a)		B	Error típ.	Wald	gl	Sig.	Exp(B)	Intervalo de confianza al 95% para Exp(B)	
								Límite inferior	Límite superior
NORMAL	Intersección	-,065	,261	,061	1	,804			
	[Sexo=F]	-,003	,300	,000	1	,993	,997	,554	1,795
	[Sexo=M]	0(b)	.	.	0	.	.	.	.
	[Droga=drogaA]	-,021	,469	,002	1	,964	,979	,391	2,453
	[Droga=drogaB]	,066	,544	,015	1	,904	1,068	,368	3,100
	[Droga=drogaC]	-21,139	,000	.	1	.	6,60E-010	6,60E-010	6,60E-010
	[Droga=drogaX]	,59[illegible]	,351	2,883	1	,090	1,816	,912	3,615
	[Droga=drogaY]	0(b)	.	.	0	.	.	.	.

a. La categoría de referencia es: ALTO

b. Este parámetro se ha establecido a cero porque es redundante

Figura 5-48

Se observa que el modelo es significativo en su conjunto (p-valor nulo del contraste de la razón de verosimilitud), pero los pseudo R-cuadrados son muy pequeños y los p-valores de los parámetros estimados del modelo son muy altos (columna *Sig,*), lo que invalida el modelo. Por otra parte, los ods ratio o razones de ventajas (columna *Exp(B)*) son bastante pequeños (solo hay uno claramente mayor que la unidad). El modelo no resulta válido y sería lógico estimar un modelo paso a paso que seleccionara las variables significativas.

El nodo Regresión Logística: Regresión logística paso a paso

El nodo *Regresión Logística* de la categoría *Clasificación* de la ficha *Modelado* permite crear y ajustar modelos de regresión logística paso a paso con la finalidad de obtener automáticamente un modelo con todas sus variables significativas. Como ejemplo vamos a ajustar un modelo logístico que estime la probabilidad de tener colesterol en función del resto de variables de entrada del ejemplo expuesto anteriormente en la regresión logística.

Comenzamos utilizando el nodo *Archivo variable* de la categoría *Orígenes* para situar como origen de datos el fichero *drug1n.txt*. A continuación se añade el nodo *Tipo* de la categoría *Operaciones con campos* al diagrama y se sitúan como campos de entrada *Sexo* y *Droga*, y como campo objetivo *Colesterol* (Figura 5-38). Para realizar el análisis de regresión logística añadimos el nodo *Logística* de la categoría *Clasificación* de la ficha *Modelado* al diagrama que ya contiene el origen de datos *drug1n.txt*. A continuación hacemos clic con el botón derecho del ratón sobre el nodo *Logística* y elegimos *Edición* en el menú emergente resultante. Se obtiene la pantalla de la regresión logística. En la solapa *Campos* elegimos *Utilizar configuración del nodo tipo* (Figura 5-49). En la solapa *Modelo* elegimos *Automático*, el método *Paso a paso hacia atrás*, regresión logística *Multinomial*, modelo con solo *Efectos principales* e *Incluir constante en ecuación* (Figura 5-50). En la solapa *Experto* en el campo *Modo* elegimos *Experto*. En la solapa *Analizar* elegimos *Calcular importancia del predictor* en el campo *Evaluación del modelo* y *Calcular puntuaciones brutas de propensión* en el campo *Puntuaciones de propensión*. En el campo *Anotaciones* elegimos *Automático* y señalamos aquellas opciones de salida que nos interesen. Al pulsar *Aceptar* se memorizan las características del análisis de la regresión logística y al hacer clic en *Ejecutar* en la parte inferior de la pantalla *Colesterol*, se realiza la regresión logística y aparece el modelo de la regresión logística en la ruta etiquetado con la variable dependiente (objetivo) del modelo (Figura 5-51).

Figura 5-49

Figura 5-50

©Alfaomega-RC Libros

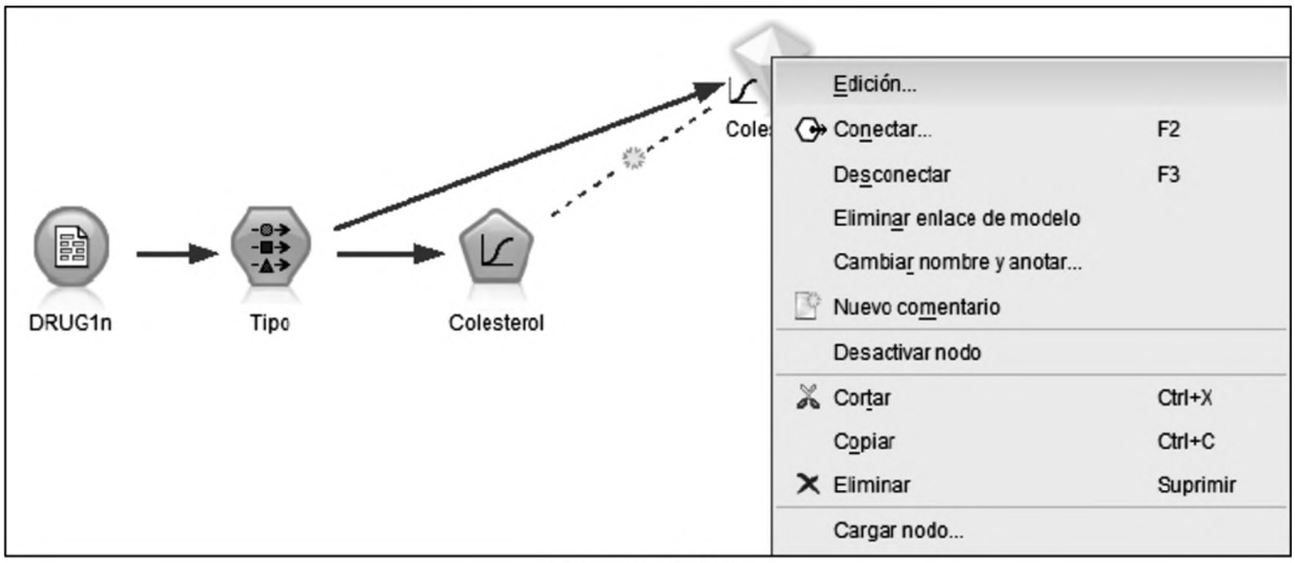


Figura 5-51

Si hacemos clic con el botón derecho del ratón sobre el modelo *Regresión* y elegimos *Edición* en el menú emergente resultante (Figura 5-51) obtenemos la pantalla de resultados de la regresión logística (Figura 5-52). En la solapa *Modelo* vemos la ecuación del modelo ajustado y la importancia de los predictores sobre la variable objetivo (Figura 5-52). En la solapa *Resumen* vemos un resumen del modelo (Figura 5-53). En la solapa *Avanzado* vemos el ajuste del modelo (Figura 5-54).

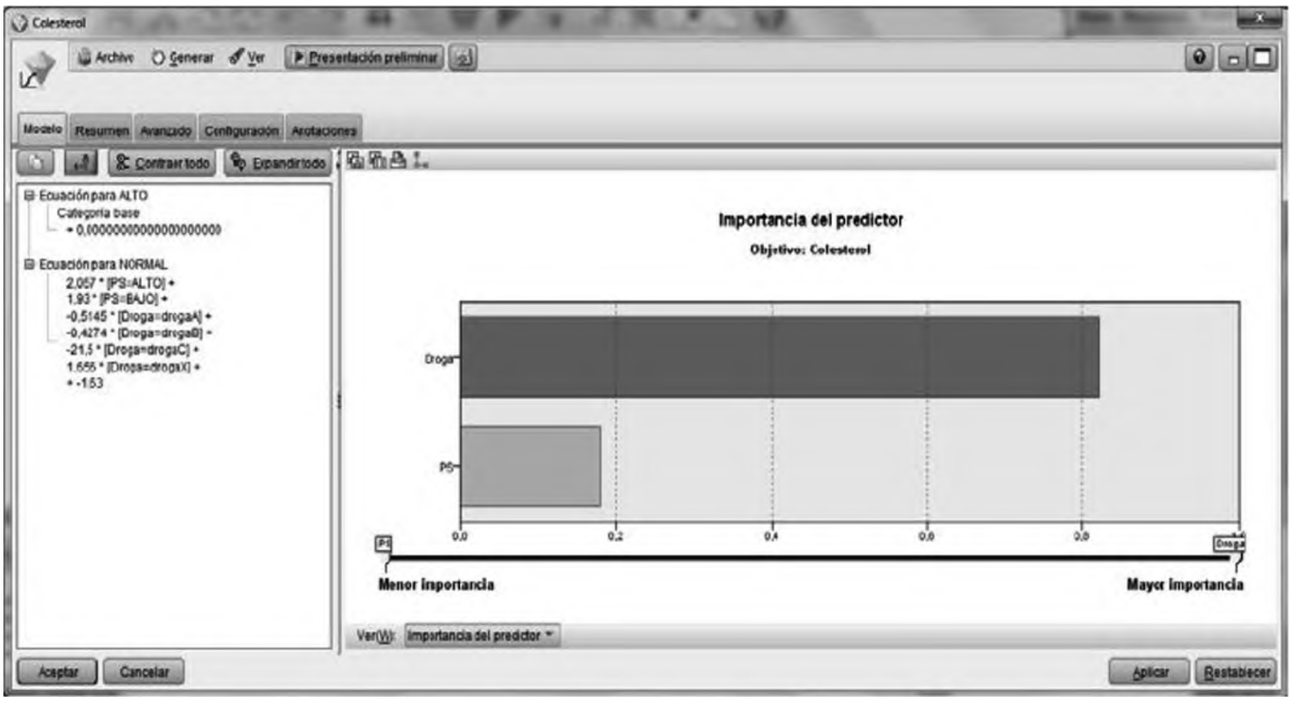


Figura 5-52

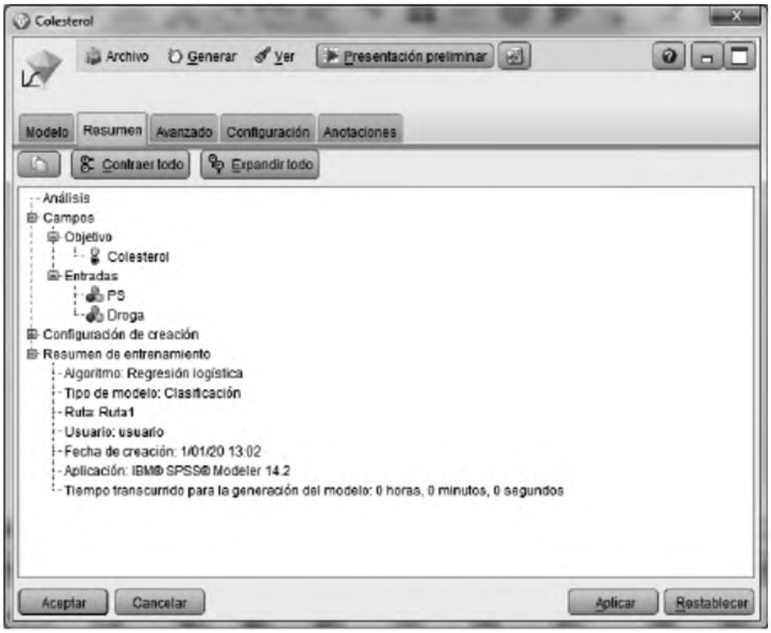


Figura 5-53

Información del ajuste del modelo

Modelo	Criterio de ajuste del modelo			Contrastes de la razón de verosimilitud		
	AIC	BIC	-2 log verosimilitud	Chi-cuadrado	gl	Sig.
Sólo la intersección	279,079	282,377	277,079			
Final	243,835	266,923	229,835	47,244	6	,000

Pseudo R cuadrado

Cox y Snell	,210
Nagelkerke	,281
McFadden	,171

Estimaciones de los parámetros

Colesterol(a)		B	Error típ.	Wald	gl	Sig.	Exp(B)	Intervalo de confianza al 95% para Exp(B)	
								Límite inferior	Límite superior
NORMAL	Intersección	-1,630	,473	11,870	1	,001			
	[PS=ALTO]	2,057	,578	12,675	1	,000	7,825	2,521	24,286
	[PS=BAJO]	1,930	,512	14,186	1	,000	6,887	2,523	18,798
	[PS=NORMAL]	0(b)	.	.	0	.	.	.	.
	[Droga=drogaA]	-,514	,533	,931	1	,335	,598	,210	1,700
	[Droga=drogaB]	-,427	,600	,507	1	,476	,652	,201	2,114
	[Droga=drogaC]	-21,504	,000	.	1	.	4,58E-010	4,58E-010	4,58E-010
	[Droga=drogaX]	1,656	,508	10,627	1	,001	5,236	1,935	14,168
	[Droga=drogaY]	0(b)	.	.	0	.	.	.	.

a. La categoría de referencia es: ALTO.

b. Este parámetro se ha establecido a cero porque es redundante.

Figura 5-54

Se observa que el modelo es significativo en su conjunto (p-valor nulo del contraste de la razón de verosimilitud). Los pseudo R-cuadrados han crecido, pero todavía siguen siendo pequeños, aunque esto no es determinane en la diagnosis.

©Alfaomega-RC Libros

Los p-valores de los parámetros estimados del modelo son muy aceptables (columna *Sig*. con valores pequeños), lo que valida el modelo. Por otra parte, los dos ratios o razones de ventajas (columna *Exp(B)*) son bastante aceptables (hay varios claramente mayores que la unidad). El modelo estimado paso a paso que ha seleccionado las variables significativas nos dice que las variables que influyen en la probabilidad de tener colesterol son la presión sanguínea y la droga.

El nodo GenLin: Modelo Lineal General

El nodo *Modelo Lineal General* de la categoría *Clasificación* de la ficha *Modelado* permite crear y ajustar modelos de regresión generales con la finalidad de incluir en los mismos todo tipo de variables realizando una diagnosis completa que valide los modelos con criterio. Como ejemplo se consideran el conjunto de datos con formato Statistics de nombre *coches.sav*. Se trata de ajustar un modelo lineal general que prediga el consumo de los automóviles (*Consumo*) en función de la cilindrada (*Motor*), el peso (*Peso*), la potencia (*Cv*) y la capacidad de aceleración (*acel*).

Comenzamos utilizando el nodo *Archivo Statatistics* de la categoría *Orígenes* para situar como origen de datos el fichero *coches.sav*. A continuación se añade el nodo *Tipo* de la categoría *Operaciones con campos* al diagrama y se sitúa como campo objetivo *Consumo* y como campos de entrada el resto de las variables del fichero (Figura 5-55). Para ajustar el modelo lineal general añadimos el nodo *GenLin* de la categoría *Clasificación* de la ficha *Modelado* al diagrama que ya contiene el origen de datos *coches.sav*. A continuación hacemos clic con el botón derecho del ratón sobre el nodo *GenLin* y elegimos *Edición* en el menú emergente resultante. Se obtiene la pantalla de la Figura 5-56.

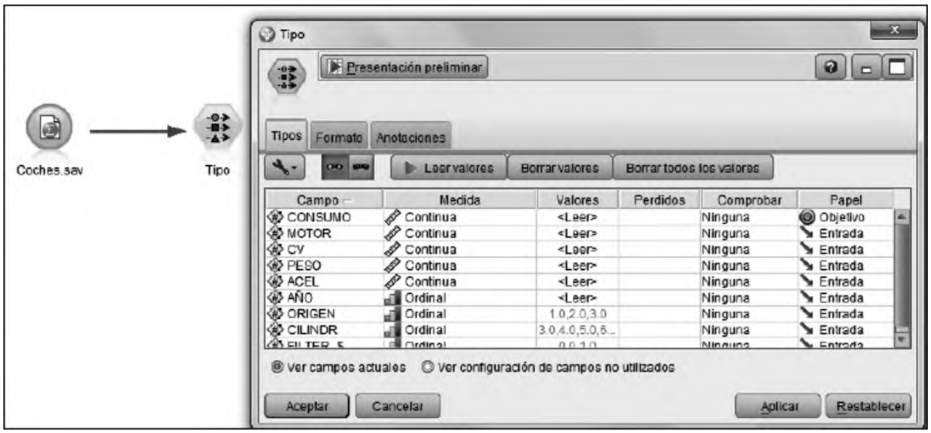


Figura 5-55

En la solapa *Campos* (Figura 5-56) elegimos *Utilizar configuración predeterminada* y elegimos las variables *Objetivo* y de *Entrada*. En la solapa *Modelo* elegimos modelo con solo *Efectos principales* e *Incluir constante en ecuación* (Figura 5-57). En la solapa *Experto* en el campo *Modo* elegimos *Experto* (Figura 5-58). En la solapa *Analizar* elegimos *Calcular importancia del predictor* en el campo *Evaluación del modelo* y *Calcular puntuaciones brutas de propensión* en el campo *Puntuaciones de propensión*. En el campo *Anotaciones* elegimos *Automático* y señalamos aquellas opciones de salida que nos interesen. Al pulsar *Aceptar* se memorizan las características del análisis de la regresión y al hacer clic en *Ejecutar* en la parte inferior de la pantalla, se realiza la regresión y aparece el modelo de la regresión lineal general en la ruta etiquetado con la variable dependiente (objetivo) del modelo (Figura 5-59).

Figura 5-56

Figura 5-57

©Alfaomega-RC Libros

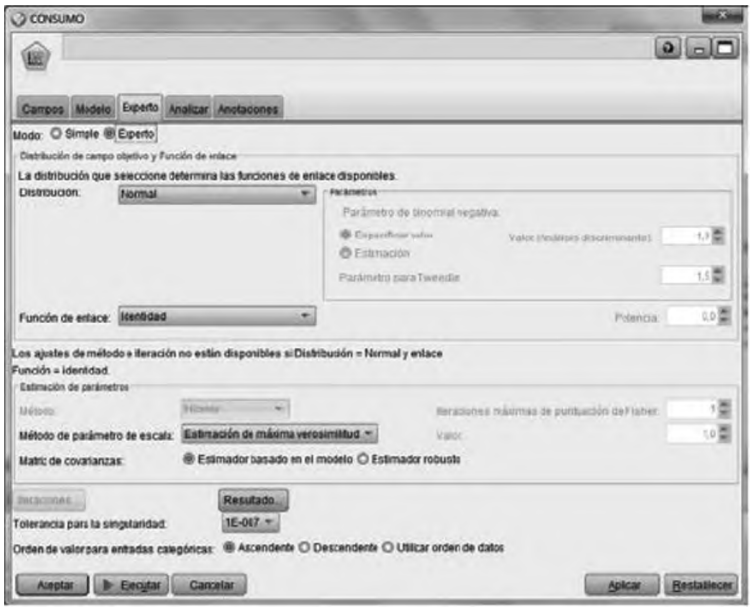


Figura 5-58

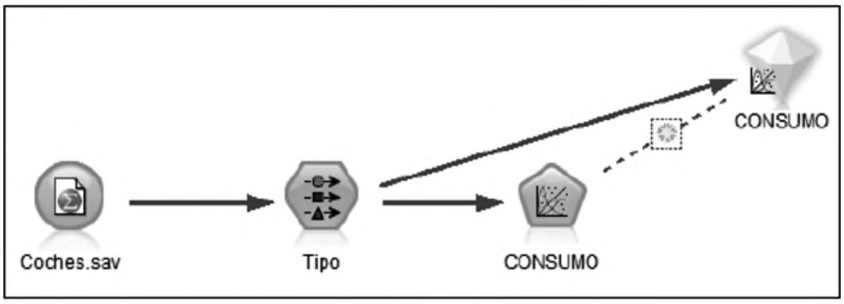


Figura 5-59

Si hacemos clic con el botón derecho del ratón sobre el modelo *GenLin* (etiquetado con la variable objetivo CONSUMO) y elegimos *Edición* en el menú emergente resultante (Figura 5-60), obtenemos la pantalla de resultados de la regresión lineal general (Figura 5-61). En la solapa *Modelo* vemos la ecuación del modelo ajustado y la importancia de los predictores sobre la variable objetivo. En la solapa *Resumen* vemos un resumen del modelo (Figura 5-62). En la solapa *Avanzado* vemos el ajuste del modelo (Figura 5-63).

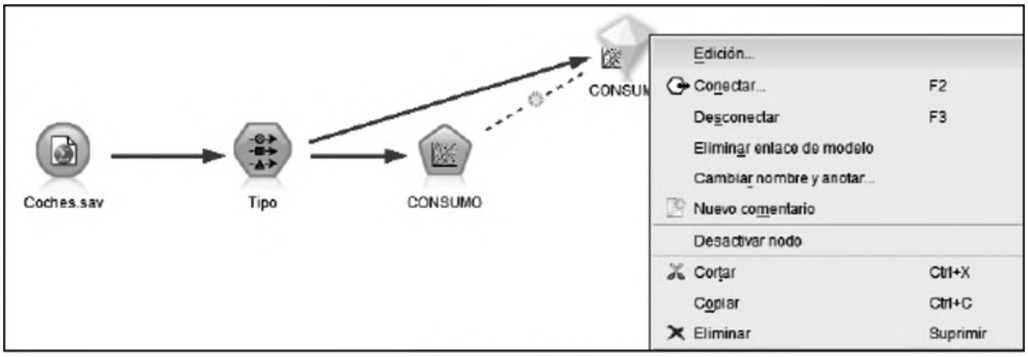


Figura 5-60

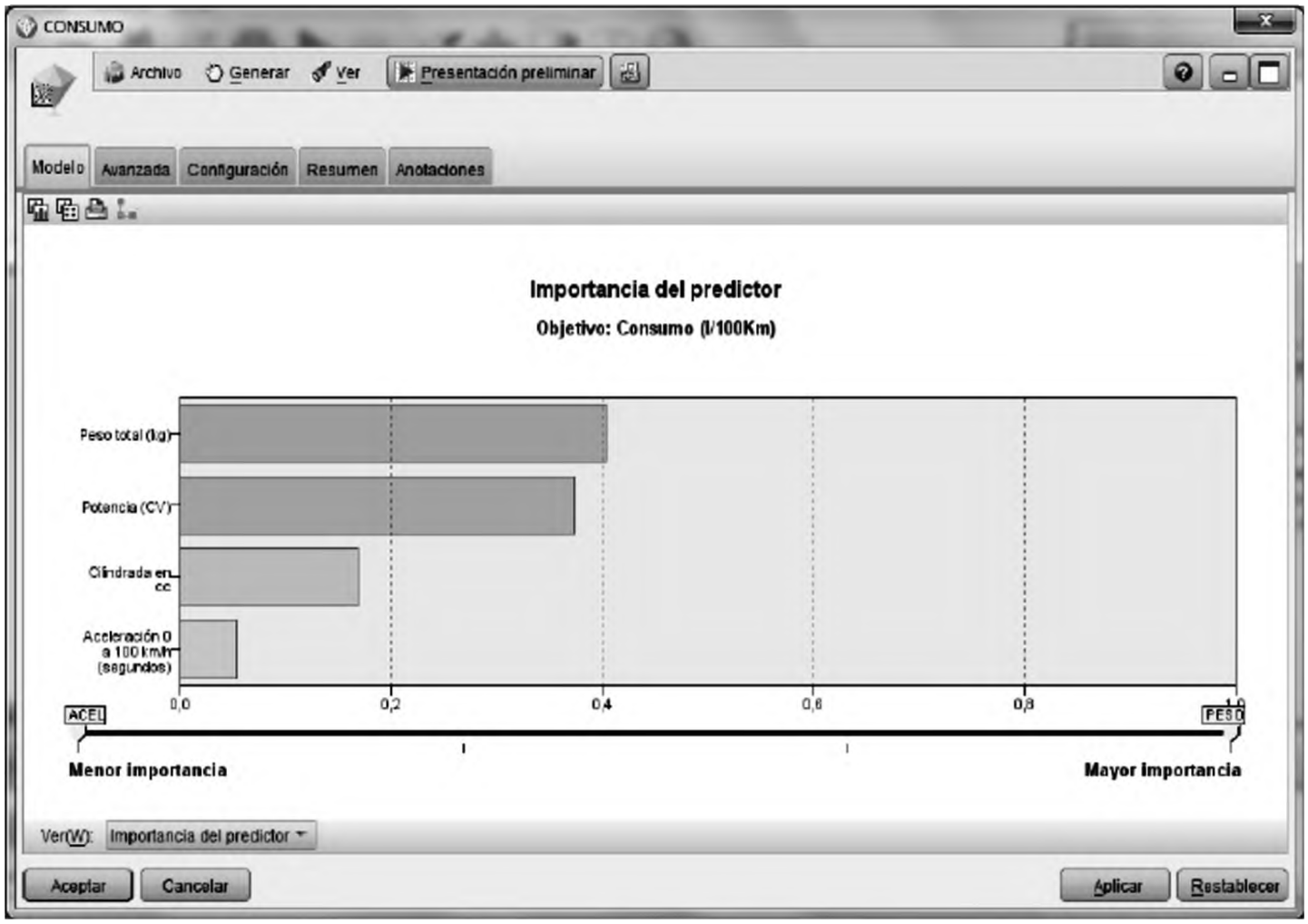


Figura 5-61

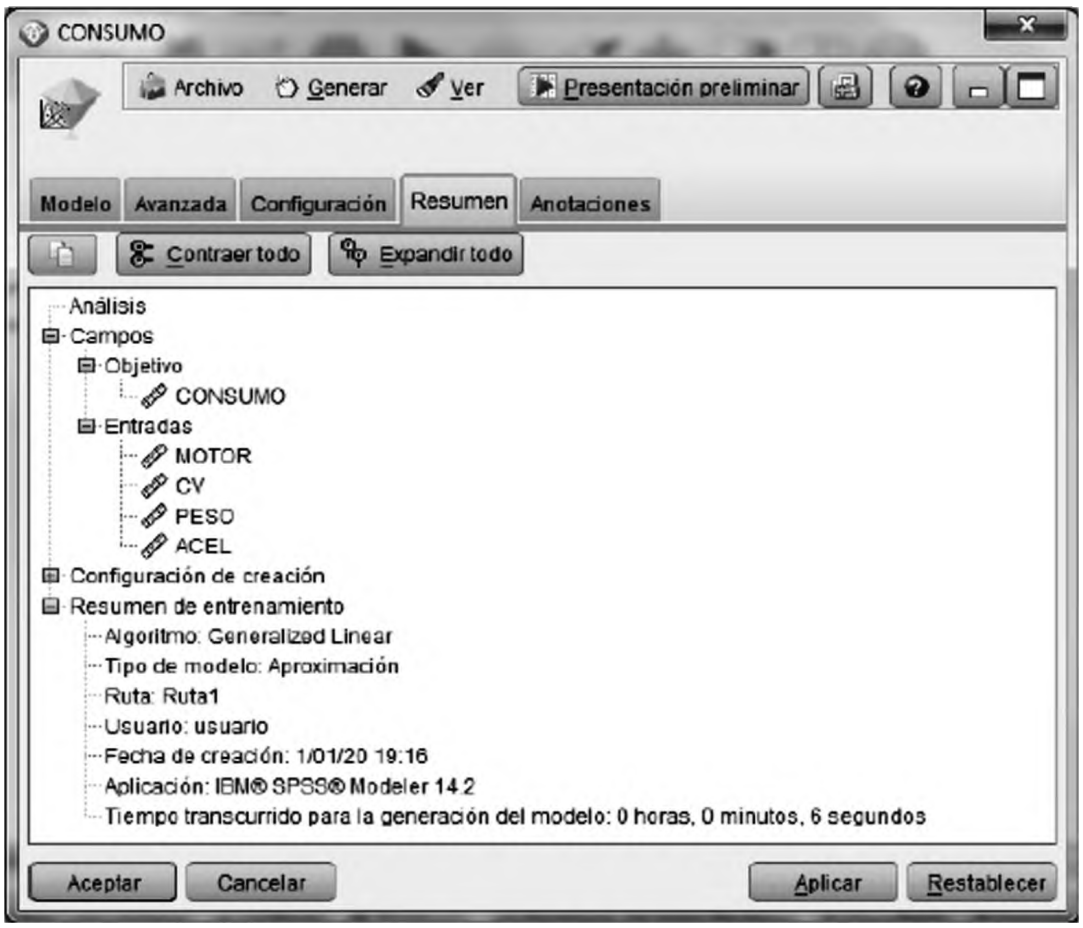


Figura 5-62

©Alfaomega-RC Libros

Contraste Omnibus(a)

Chi-cuadrado de la razón de verosimilitudes	gl	Sig.
551,157	4	,000

Variable dependiente: Consumo (l/100Km)Modelo: (Intersección), MOTOR, CV, PESO, ACEL

a. Compara el modelo ajustado con el modelo con sólo la intersección.

Contrastes de los efectos del modelo

Origen	Tipo III		
	Chi-cuadrado de Wald	gl	Sig.
(Intersección)	,139	1	,709
MOTOR	2,634	1	,105
CV	31,558	1	,000
PESO	19,641	1	,000
ACEL	,182	1	,669

Variable dependiente: Consumo (l/100Km)Modelo: (Intersección), MOTOR, CV, PESO, ACEL

Estimaciones de los parámetros

Parámetro	B	Tip. Error	Intervalo de confianza de Wald 95%		Contraste de hipótesis		
			Inferior	Superior	Chi-cuadrado de Wald	gl	Sig.
(Intersección)	,432	1,1590	-1,840	2,704	,139	1	,709
MOTOR	,000	,0002	-6,42E-005	,001	2,634	1	,105
CV	,044	,0078	,029	,059	31,558	1	,000
PESO	,005	,0011	,003	,007	19,641	1	,000
ACEL	,025	,0587	-,090	,140	,182	1	,669
(Escala)	3,832(a)	,2737	3,331	4,408			

Variable dependiente: Consumo (l/100Km)Modelo: (Intersección), MOTOR, CV, PESO, ACEL

a. Estimación de máxima verosimilitud.

Figura 5-63

Se observa que el modelo es significativo en su conjunto (p-valor cero en el contraste de la chi-cuadrado de la razón de verosimilitudes) y que los parámetros son significativos individualemnte salvo para la variable *Acel* y la constante del modelo. En este caso se comienza quitando la constante del modelo en la solapa *Modelo* de la regresión (Figura 5-64). Volviendo a ejecutar se solucionan los problemas de no significatividad individual de los parámetros (Figura 5-65).

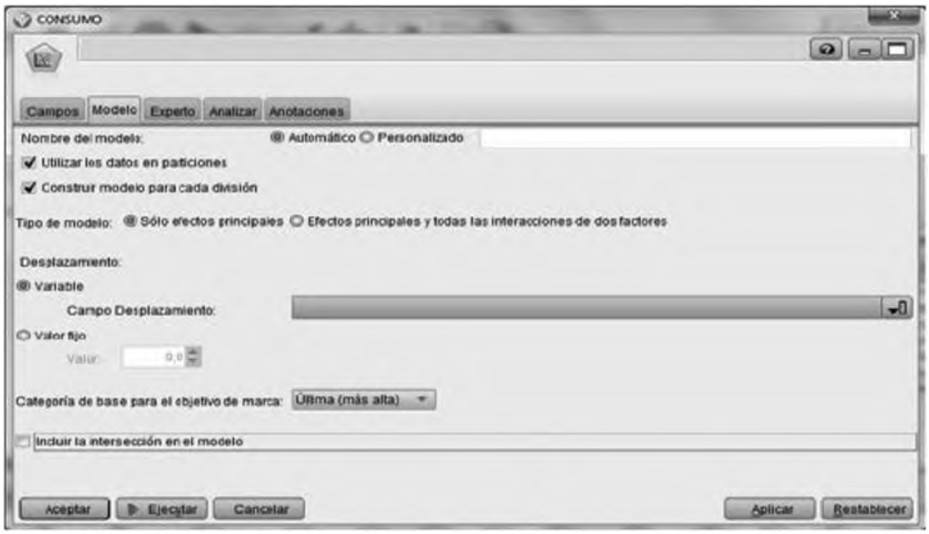


Figura 5-64

Contraste Omnibus(a)

Chi-cuadrado de la razón de verosimilitudes	gl	Sig.
1417,416	4	,000

Variable dependiente: Consumo (l/100Km)Modelo: MOTOR, CV, PESO, ACEL

a. Compara el modelo ajustado con el modelo nulo.

Contrastes de los efectos del modelo

Origen	Tipo III		
	Chi-cuadrado de Wald	gl	Sig.
MOTOR	2,494	1	,114
CV	66,287	1	,000
PESO	19,501	1	,000
ACEL	2,448	1	,118

Variable dependiente: Consumo (l/100Km)Modelo: MOTOR, CV, PESO, ACEL

Estimaciones de los parámetros

Parámetro	B	Tip. Error	Intervalo de confianza de Wald 95%		Contraste de hipótesis		
			Inferior	Superior	Chi-cuadrado de Wald	gl	Sig.
MOTOR	,000	,0002	-7,07E-005	,001	2,494	1	,114
CV	,046	,0056	,035	,057	66,287	1	,000
PESO	,005	,0011	,003	,007	19,501	1	,000
ACEL	,044	,0283	-,011	,100	2,448	1	,118
(Escala)	3,833(a)	,2738	3,333	4,410			

Variable dependiente: Consumo (l/100Km)Modelo: MOTOR, CV, PESO, ACEL

a. Estimación de máxima verosimilitud.

Figura 5-65

Si en la pantala de la regresión en la solapa *Modelo* elegimos *Efectos principales* y todas las interacciones de dos factores (Figura 5-66) se obtiene el modelo lineal generalizado ajustado de las Figuras 5-67 y 5-68.

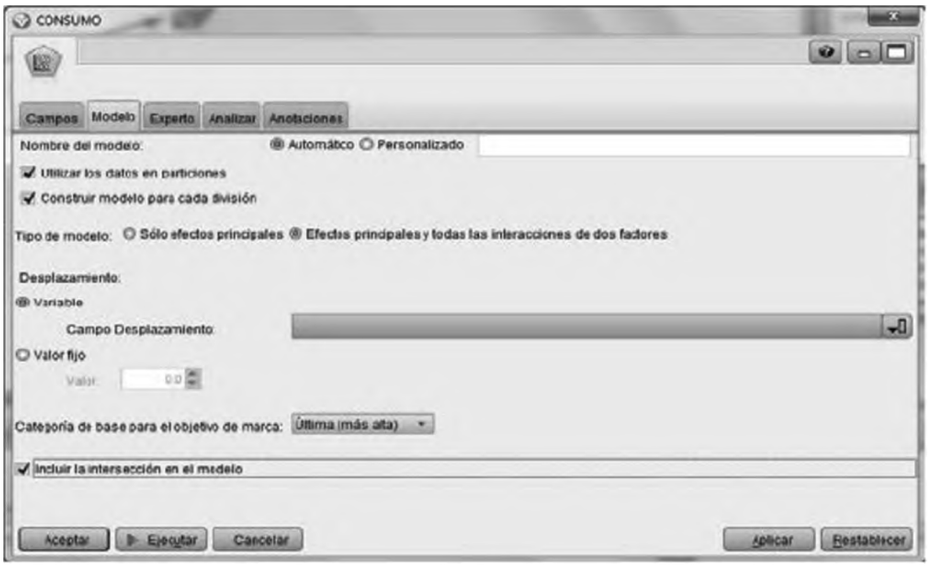


Figura 5-66

Contraste Omnibus(a)

Chi-cuadrado de la razón de verosimilitudes	gl	Sig.
650,635	10	,000

Variable dependiente: Consumo (l/100Km)Modelo: (Intersección), MOTOR, MOTOR * CV, MOTOR * PESO, MOTOR * ACEL, CV, CV * PESO, CV * ACEL, PESO, PESO * ACEL, ACEL

a. Compara el modelo ajustado con el modelo con sólo la intersección.

Figura 5-67

©Alfaomega-RC Libros

Parámetro	B	Ejp. Error	Intervalo de confianza de Wald 95%		Contraste de hipótesis		
			Inferior	Superior	Chi-cuadrado de Wald	gl	Sig.
(Intersección)	19,333	4,8274	9,872	28,795	16,040	1	,000
MOTOR	.004	.0016	.001	.008	7,430	1	.006
MOTOR * CV	-3,32E-005	5,251E-006	-4,35E-005	-2,29E-005	39,970	1	,000
MOTOR * PESO	2,46E-006	4,877E-007	1,51E-006	3,42E-006	25,520	1	,000
MOTOR * ACEL	,001	7,198E-005	[illegible]	-1,64E-005	4,785	1	,029
CV	,151	,0409	[illegible]	,231	13,592	1	,000
CV * PESO	,000	3,297E-005	3,55E-005	,000	9,216	1	,002
CV * ACEL	,004	,0020	-,008	-,001	5,036	1	,025
PESO	-,049	,0051	-,065	-,032	34,054	1	,000
PESO * ACEL	,002	,0004	,001	,003	27,415	1	,000
ACEL	-,781	,2250	-1,164	[illegible]	9,884	1	,002
(Escala)	2,973(a)	,2124	2,585	3,420			

Variable dependiente: Consumo (l/100Km)Modelo: (Intersección), MOTOR, MOTOR * CV, MOTOR * PESO, MOTOR * ACEL, CV, CV * PESO, CV * ACEL, PESO, PESO * ACEL, ACEL

a. Estimación de máxima verosimilitud.

Figura 5-68

La Figura 5-67 muestra la significatividad conjunta de los parámetros estimados del modelo (p-valor muy pequeño del contraste de la chi-cuadrado de la razón de verosimilitudes) y la significatividad individual de los parámetros estimados del modelo incluyendo la constante y las interacciones de orden dos de todas sus variables predictoras (Figura 5-68).

El nodo Discriminante: Modelos de análisis discriminante

El nodo *Discriminante* de la categoría *Clasificación* de la ficha *Modelado* permite crear y ajustar modelos de análisis discriminante con la finalidad de clasificar registros. Como ejemplo se consideran el conjunto de datos de nombre *coches.sav*. Se trata de ajustar un modelo discriminante que prediga el origen de los automóviles en función de sus características de consumo, cilindrada, peso, potencia y aceleración.

Comenzamos utilizando el nodo *Archivo Statistics* de la categoría *Orígenes* para situar como origen de datos el fichero *Coches.sav*. A continuación se añade el nodo *Tipo* de la categoría *Operaciones con campos* al diagrama y se sitúa como campo objetivo *Origen* y como campos de entrada el resto de las variables del fichero (Figura 5-69). Para realizar el análisis discriminante añadimos el nodo *Discriminante* de la categoría *Clasificación* de la ficha *Modelado* al diagrama que ya contiene el origen de datos *Coches.sav*. A continuación hacemos clic con el botón derecho del ratón sobre el nodo *Discriminante* y elegimos *Edición* en el menú emergente resultante. Se obtiene la pantalla de la Figura 5-70. En la solapa *Campos* elegimos *Utilizar configuración personalizada* e introducimos la variable *Objetivo* y las variables de *Entrada* del modelo discriminante. En la solapa *Modelo* elegimos *Automático* y el método *Introducir* (Figura 5-71). En la solapa *Experto* en el campo *Modo* elegimos *Experto* (Figura 5-72) y en el botón *Resultados* elegimos la salida del discriminante (Figura 5-73). En la solapa *Analizar* elegimos *Calcular importancia del predictor* en el campo *Evaluación del modelo* y *Calcular puntuaciones brutas de propensión* en el campo *Puntuaciones de propensión*. En el campo *Anotaciones* elegimos *Automático* y señalamos aquellas opciones de salida que nos interesen.

Al pulsar *Aceptar* se memorizan las características del análisis de la regresión logística y al hacer clic en *Ejecutar* en la parte inferior de la pantalla, se realiza la regresión logística y aparece el modelo de la regresión logística en la ruta etiquetado con la variable dependiente (objetivo) del modelo (Figura 5-74).

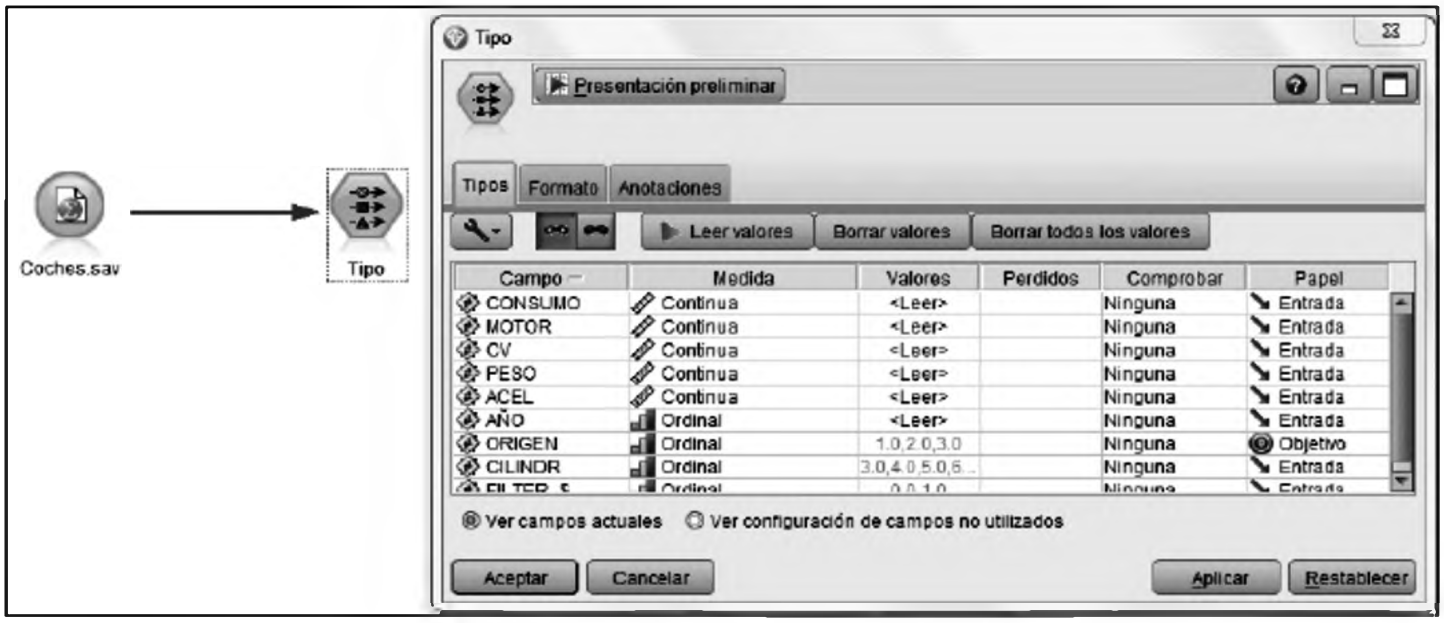


Figura 5-69

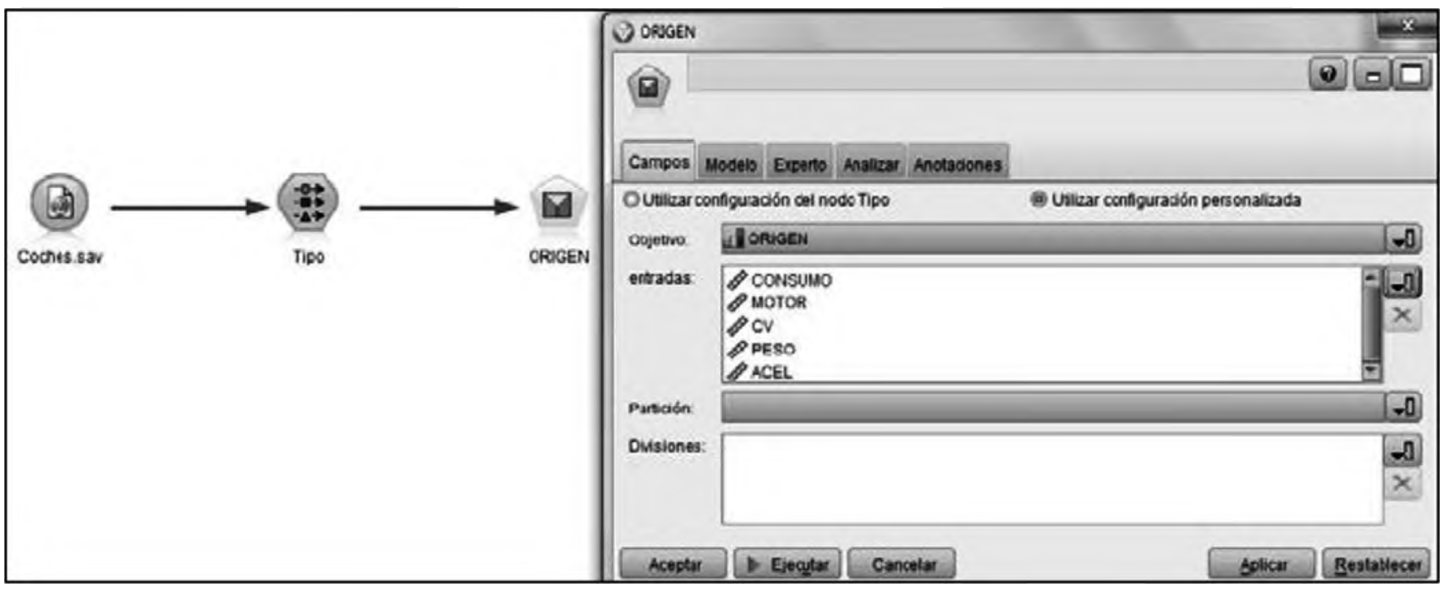


Figura 5-70

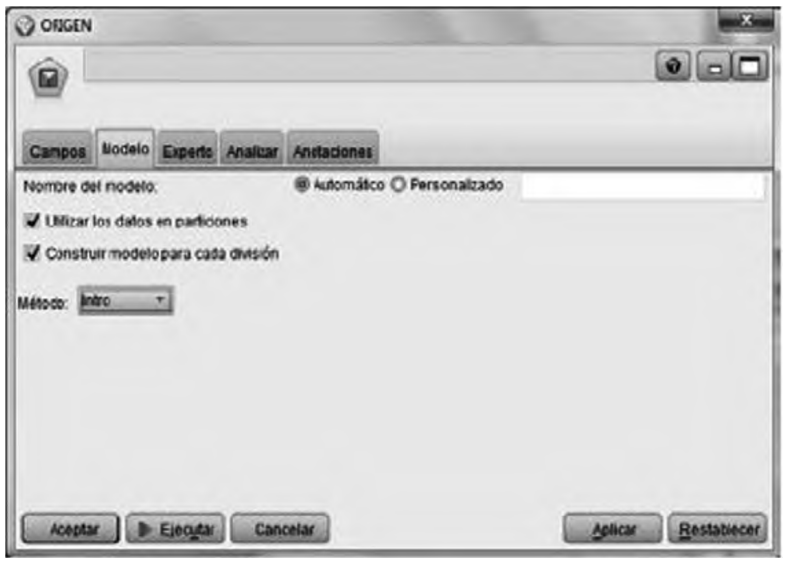


Figura 5-71

Figura 5-72

©Alfaomega-RC Libros

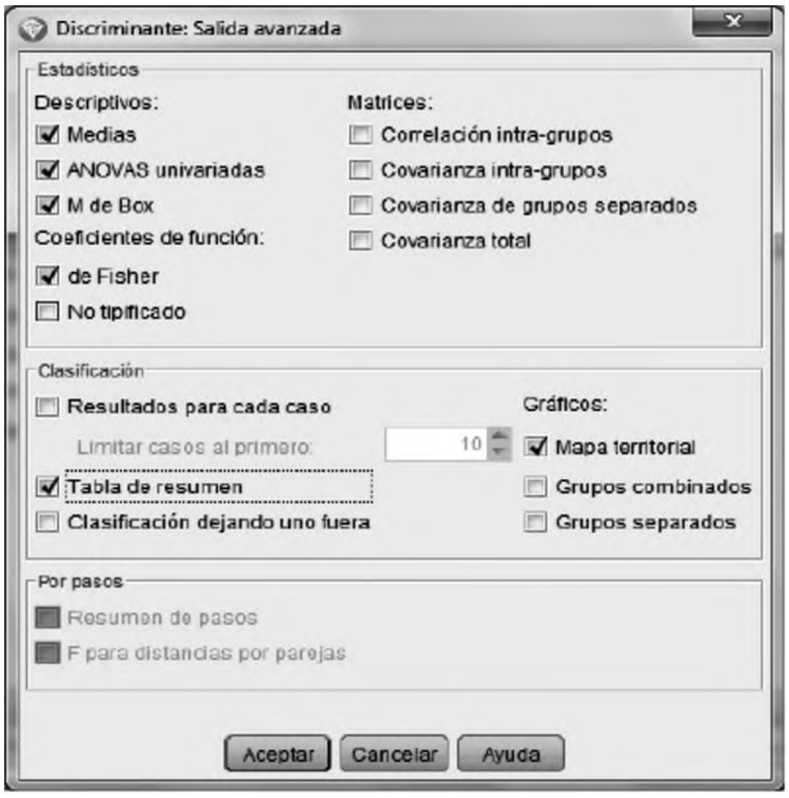


Figura 5-73

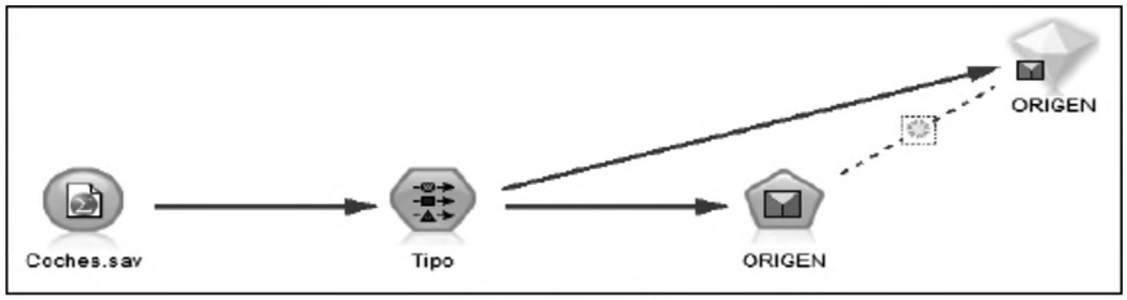


Figura 5-74

Si hacemos clic con el botón derecho del ratón sobre el modelo *Discriminante* y elegimos *Edición* en el menú emergente resultante (Figura 5-75), obtenemos la pantalla de resultados del análisis discriminante (Figura 5-76). En la solapa *Modelo* vemos la ecuación del modelo ajustado y la importancia de los predictores sobre la variable objetivo (Figura 5-46). En la solapa *Resumen* vemos un resumen del modelo y en la solapa *Avanzado* vemos el ajuste del modelo.

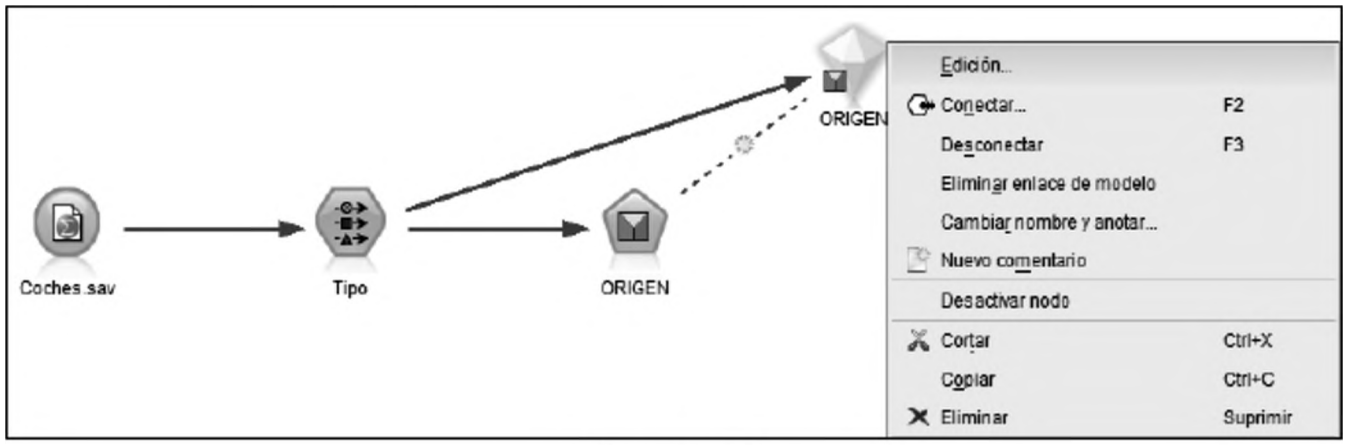


Figura 5-75

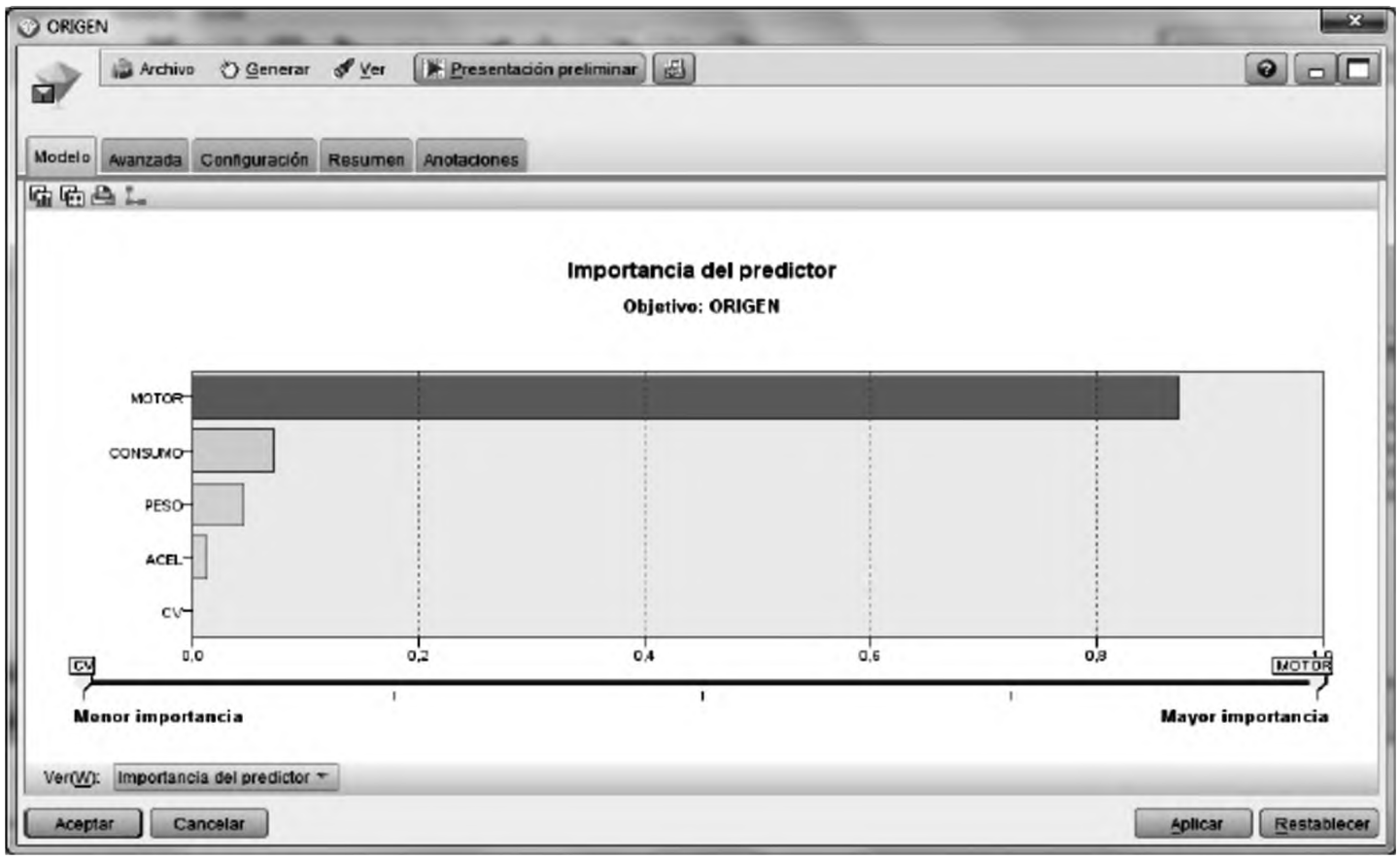


Figura 5-76

©Alfaomega-RC Libros

CAPÍTULO 6

MODELADO PREDICTIVO CON HERRAMIENTAS DE SAS

TÉCNICAS DE MODELADO PREDICTIVO CON SAS ENTERPRISE MINER

Enterprise Miner implementa las técnicas predictivas en la opción *Model* (Figura 6-1).

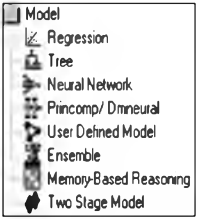


Figura 6-1

MODELO DE REGRESIÓN MÚLTIPLE CON EL NODO REGRESSION

A través del nodo *Regression* es posible realizar regresiones múltiples y regresiones logísticas. Una vez que disponemos ya de nuestros datos de trabajo en formato SAS (archivo *educa.sas7bdat*) en una determinada librería (librería *Trabajo* que representa el subdirectorio *c:\libros\miningt*), abrimos el proyecto P1 (*File → Open*) y mediante *File → New → Diagram* creamos el diagrama D11. A continuación, el nodo *Regresion* se obtiene por medio del botón *Tools* del navegador de proyecto de Enterprise Miner como subopción de la categoría *Model* o arrastrando el propio nodo sobre la zona de trabajo al lado del nodo *Input Data Source* al que se le ha asignado previamente el conjunto de datos *educa.sas7bdat* de la librería TRABAJO.

A continuación se realizará la unión de ambos (Figura 6-2).

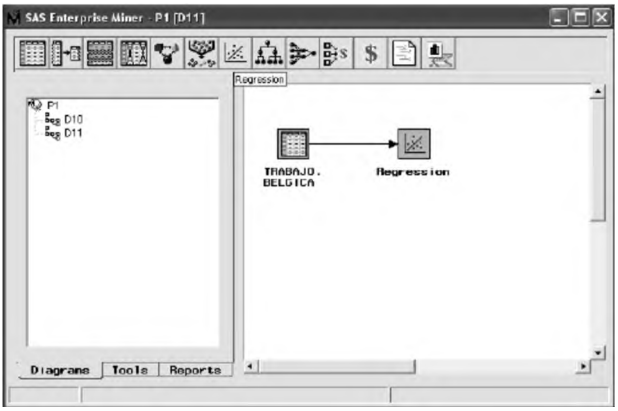


Figura 6-2

Nos planteamos un modelo de regresión que tenga la variable *Math,* que mide los resultados de las pruebas de matemáticas, como dependiente y el resto de las variables (salvo *Read* como independientes). Por lo tanto será necesario declarar las variables *Math* y *Read* como de tipo *Target* y el resto de tipo *Input*. Esta tarea se realiza haciendo doble clic sobre el nodo *Input Data Source* etiquetado como TRABAJO.BELGICA y eligiendo la pestaña *Variables*. A continuación se hace clic con el botón derecho del ratón sobre la variable *Math*, se elige *Set Model Role* (Figura 6-3) y a continuación *Target* (Figura 6-4). Se repite el proceso con la variable *Read* y ya se tienen las dos variables de tipo *Target* o dependientes (Figura 6-5). El resto son de tipo *Input* o independientes por defecto.

 ©Alfaomega-RC Libros

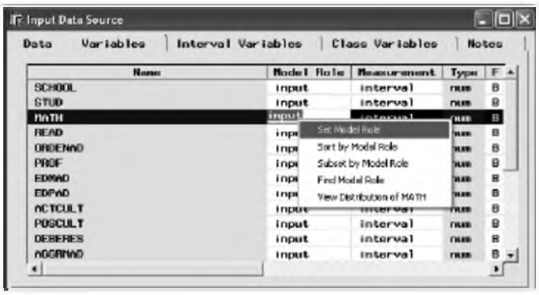


Figura 6-3

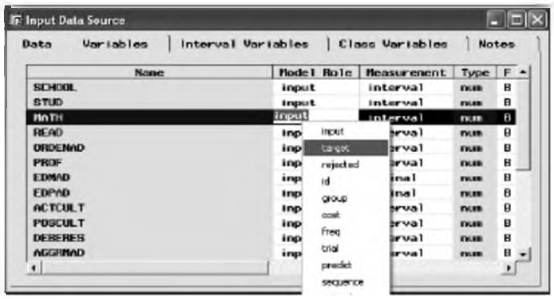


Figura 6-4

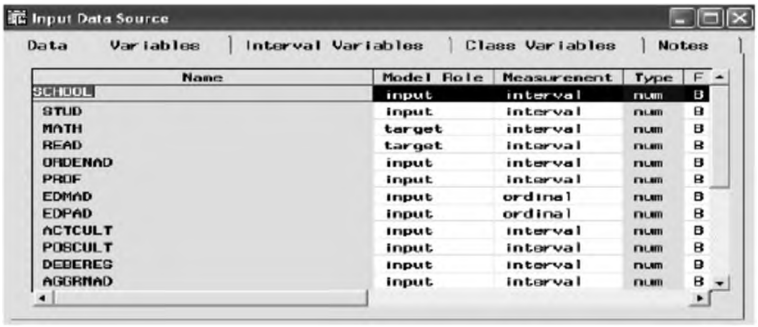


Figura 6-5

A continuación se hace doble clic sobre el nodo *Regression* y se obtiene la pantalla de selección de la variable dependiente que ofrece aquellas variables tipo *Target* definidas en el archivo de datos (Figura 6-6). Elegimos *Math* como variable dependiente del modelo y pulsamos *OK*. Se obtiene la pantalla de la Figura 6-7 en cuya pestaña *Variables* activaremos las variables explicativas o *inputs* asignándoles *Status* → *use*. Si no queremos incluir alguna de las variables, cambiaremos su *Status* a *don't use*. Para ello se hace clic con el botón derecho del ratón en la columna *Status* sobre la variable a cambiar de estado, se elige *Set Status* (Figura 6-7) y a continuación *use* o *don't use* (Figura 6-8) según se necesite.

En la pestaña *Selection Method* debemos seleccionar el método para la selección de variables explicativas (Figura 6-10). En este ejemplo hemos seleccionado la opción *None*, que implica que todas las variables explicativas serán introducidas en un único paso en el análisis.

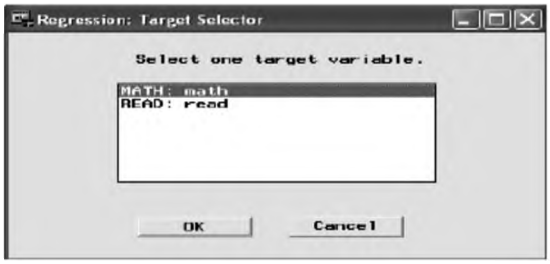


Figura 6-6

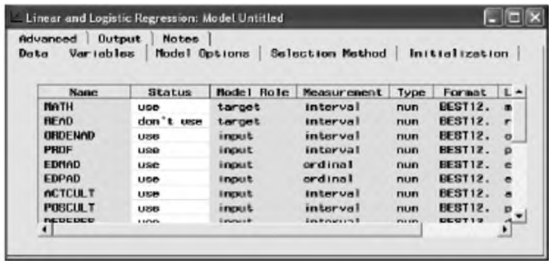


Figura 6-7

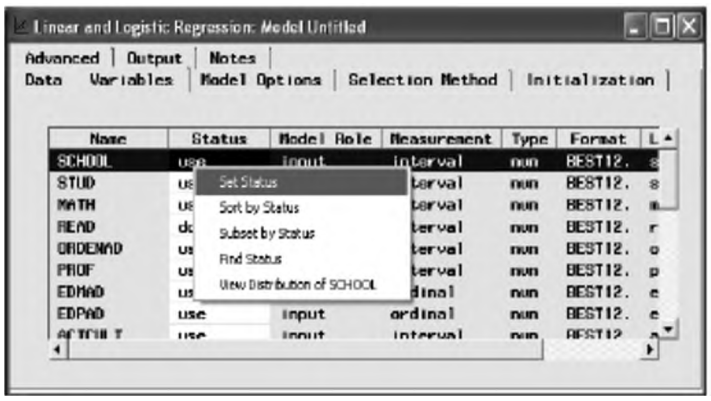


Figura 6-8

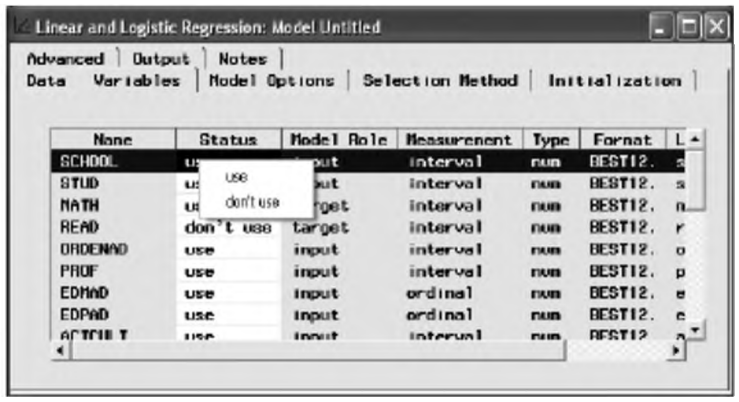


Figura 6-9

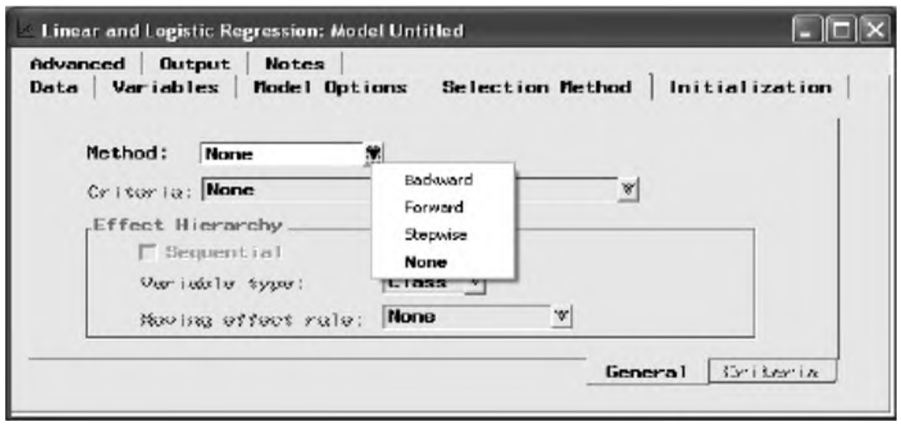


Figura 6-10

En la pestaña *Output* (subpestaña *Score Data Set*) fijamos las especificaciones acerca de dónde será guardada la información con las predicciones del modelo (Figura 6-11). En este caso las predicciones realizadas sobre los datos de entrenamiento, únicos disponibles, serán guardados en la librería *EMDATA* y dentro de esta en el archivo *STRNHFST*.

Por otro lado, en la pestaña *Output* (subpestaña *Parameter Estimates*) se muestra el destino del archivo donde quedarán guardados los parámetros estimados (Figura 6-12).

Figura 6-11

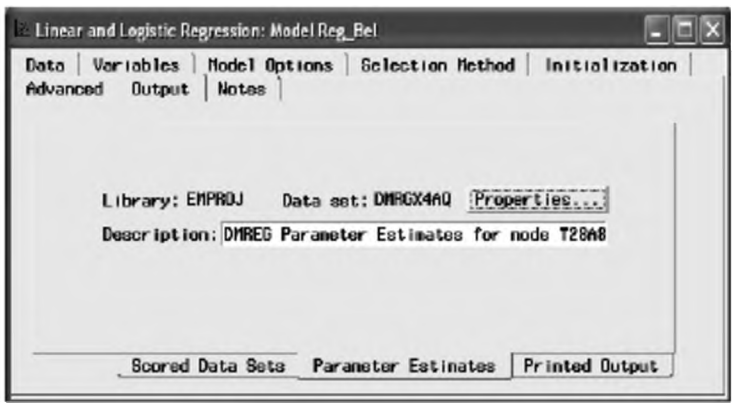


Figura 6-12

Finalmente en la ventana *Output* (subpestaña *Printed Output*), seleccionaremos los resultados que queremos obtener (Figura 6-13).

©Alfaomega-RC Libros

Una vez seleccionadas todas estas acciones estaremos en disposición de ejecutar la regresión. Previamente, al cerrar el nodo, el sistema nos pedirá que asignemos un nombre al modelo construido (Figura 6-14). Una vez nombrado y descrito pulsaremos *OK* y ejecutaremos el nodo de la forma habitual.

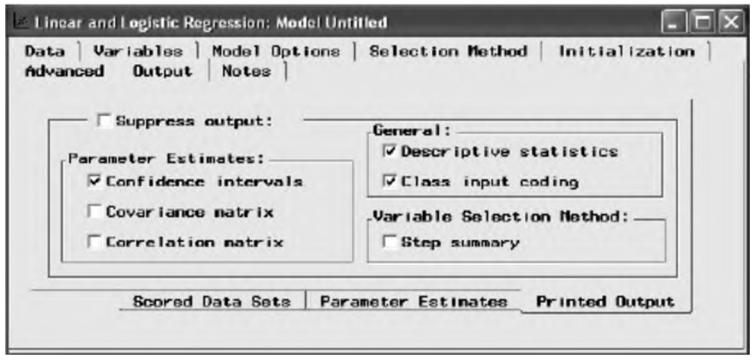


Figura 6-13

Figura 6-14

También puede ejecutarse la regresión haciendo clic con el botón derecho del ratón sobre el nodo *Regression* y eligiendo *Run* en el menú emergente resultante (Figura 6-15). A continuación aparece la Figura 6-16 cuyo botón *Yes* permite ver los resultados (Figura 6-17).

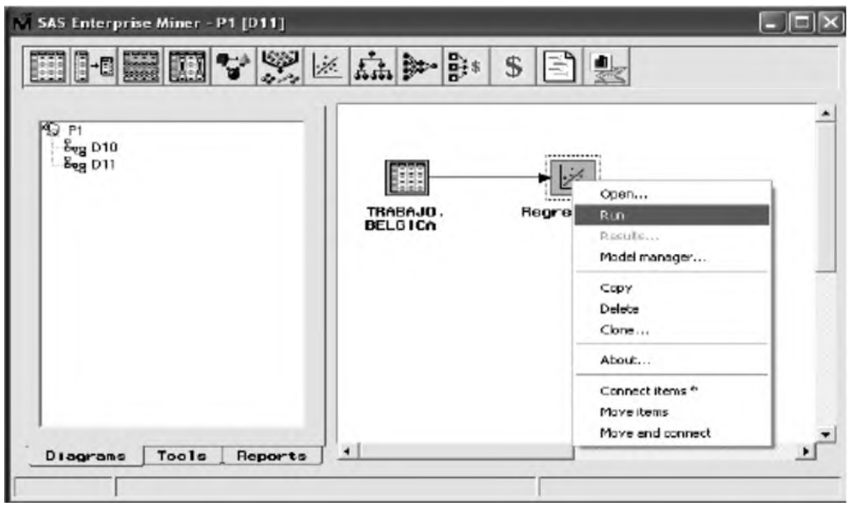


Figura 6-15

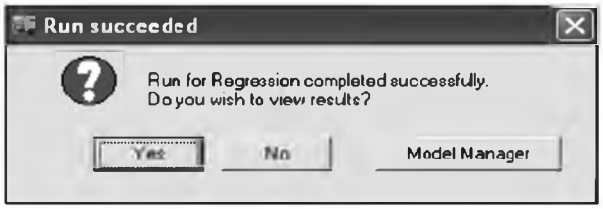


Figura 6-16

Para *visualizar los resultados de forma gráfica* podemos seleccionar en la pestaña *Estimates* la opción *Effect T-scores* o importancia de cada regresor (Figura 6-17).

Otra posibilidad es seleccionar *Estimates* o tamaño del coeficiente calculado (Figura 6-18).

Mediante la opción *Table* podemos analizar los coeficientes obtenidos para cada regresor junto a su estadístico *T* de Student (Figura 6-19).

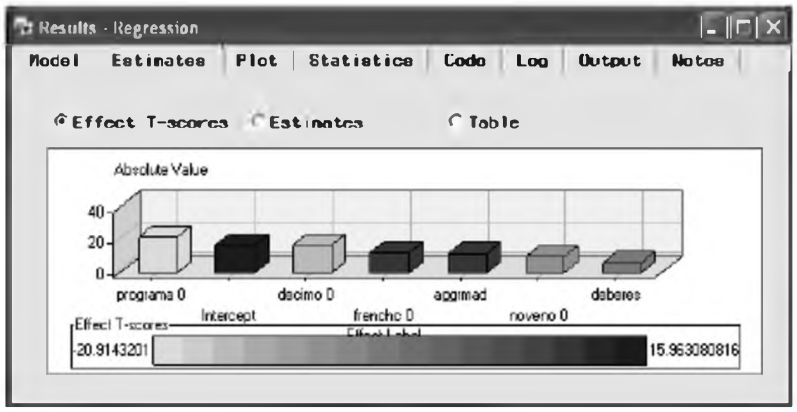


Figura 6-17

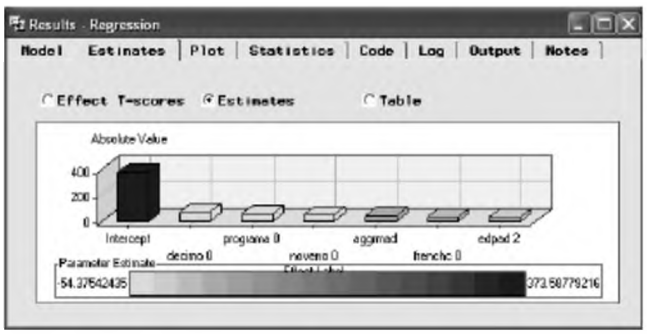


Figura 6-18

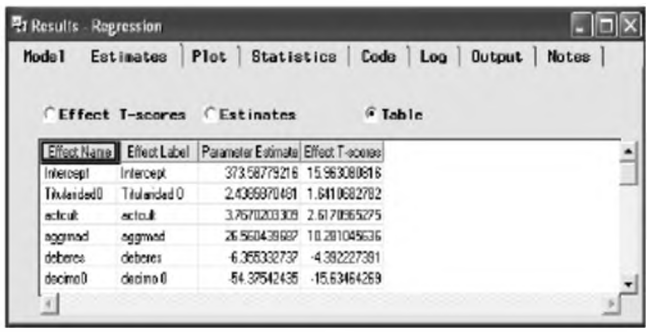


Figura 6-19

En la ventana *Plot* podemos construir distintos gráficos entre las variables y los resultados estimados (Figura 6-20). En la ventana *Statistics* (Figura 6-21) se muestran algunos estadísticos de ajuste usuales del modelo ejecutado.

©Alfaomega-RC Libros

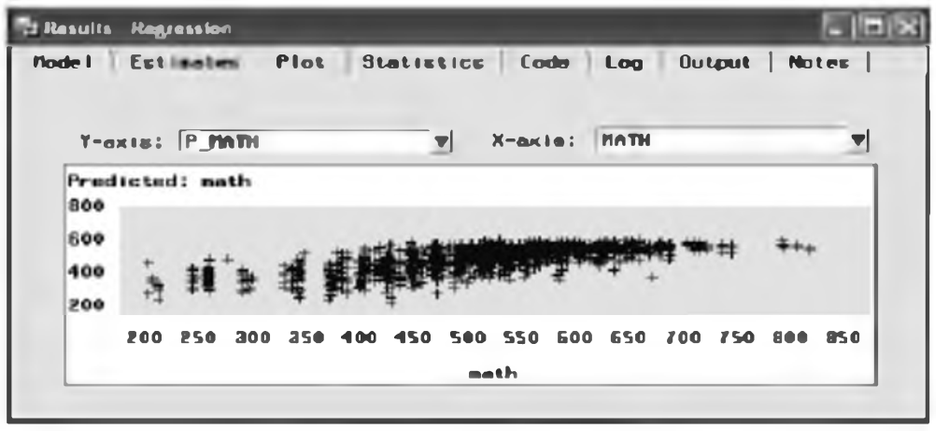


Figura 6-20

Figura 6-21

La ventana *Code* muestra el código SAS generado que permite ejecutar el análisis en el *Editor de comandos*. Finalmente, de especial importancia es la ventana *Output*. En ella se muestran todos los resultados de la regresión (Figura 6-22). Esta información comprende la codificación de las variables nominales, estadísticas descriptivas, análisis de varianza o de relevancia del modelo, coeficiente de determinación y coeficiente de determinación ajustado y otros estadísticos de ajuste. Además, comprende las estimaciones de los parámetros con *p*-valores e intervalos de confianza (Figura 6-23).

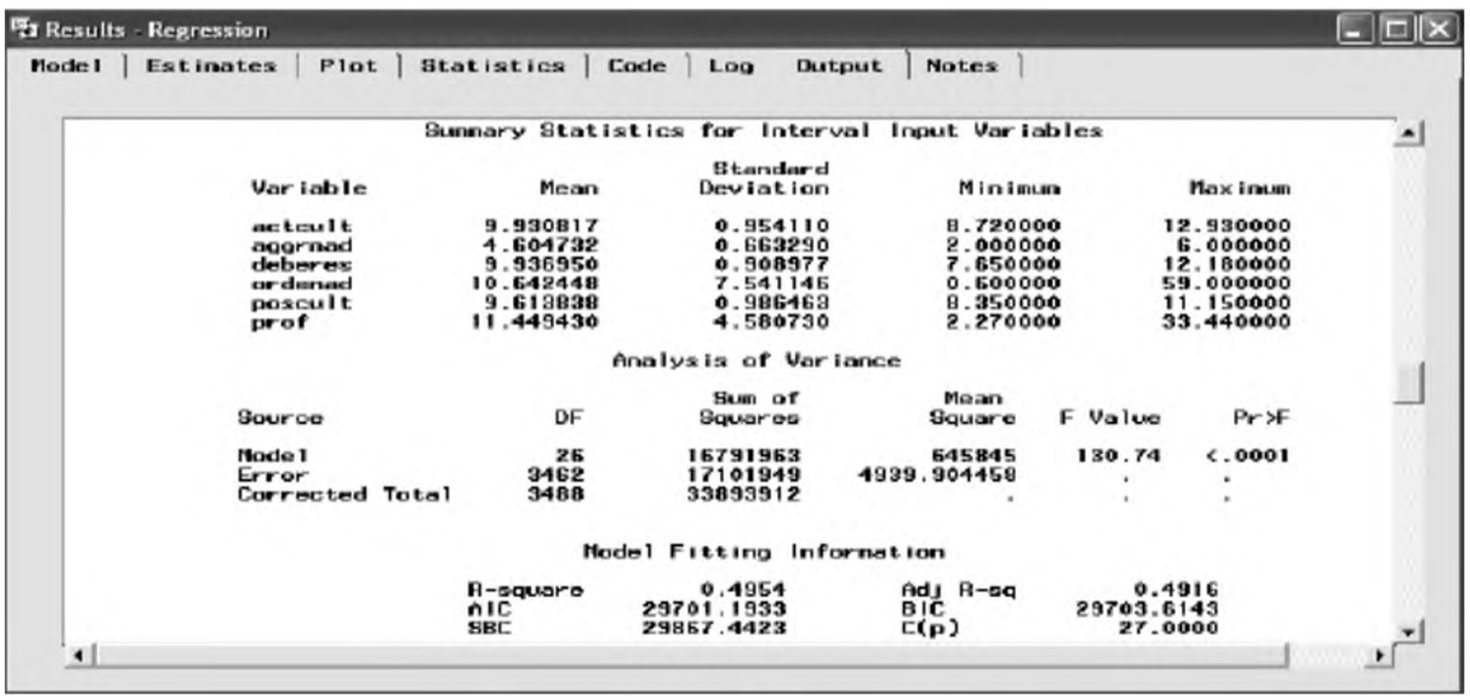


Figura 6-22

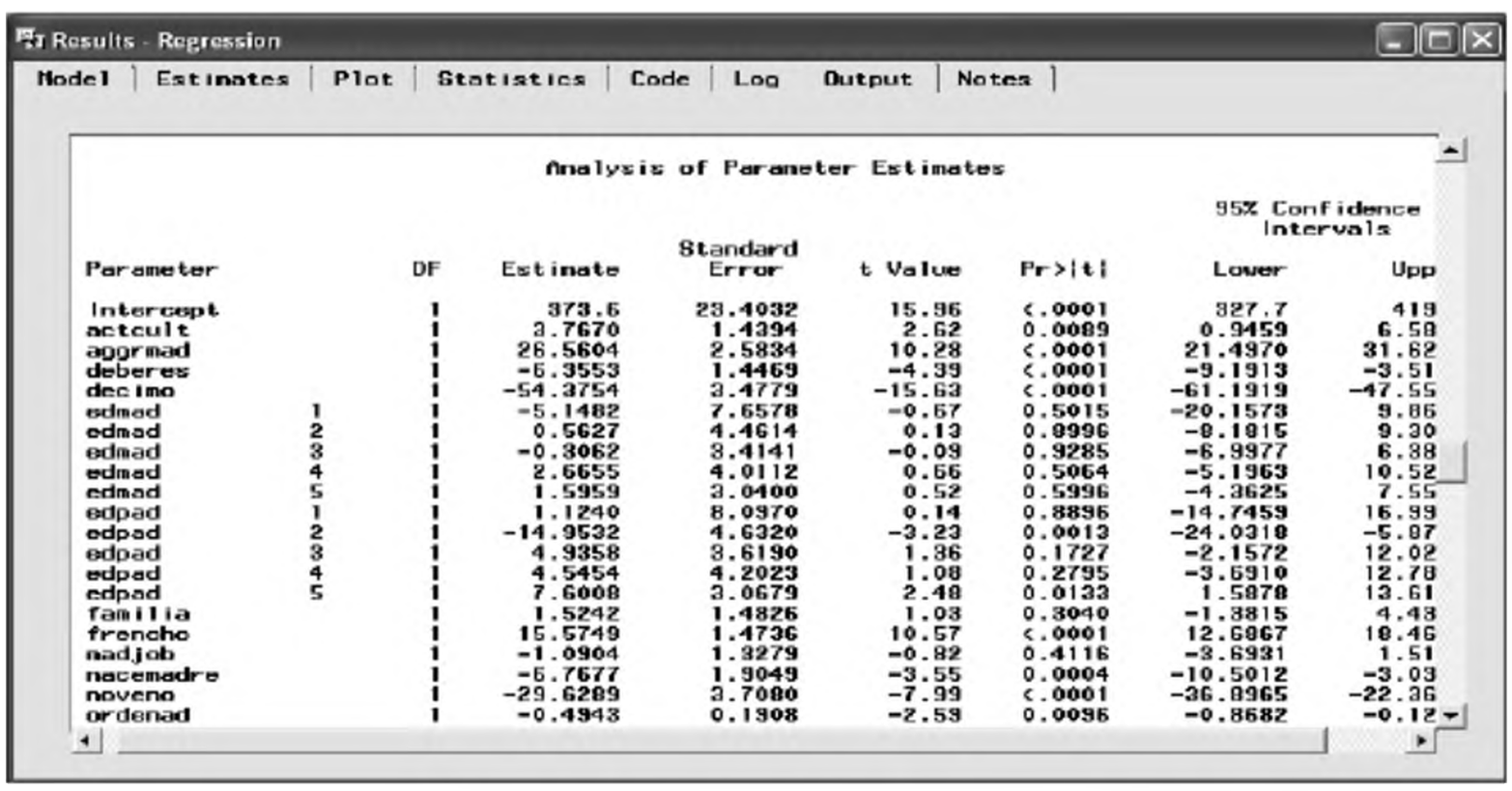


Figura 6-23

Si queremos ver o analizar las predicciones y los residuos obtenidos, estos están disponibles en el fichero creado guardado en la librería *EMDATA* y dentro de esta en el archivo *STRNHFST* (Figura 6-24).

VIEWTABLE: Scored Train:

	programa	repite	madjob	padjob	nacemadre	Titularidad	Warnings	Predicted: math	Residual: math
1	1	0	1	1	1	1		559.52637919	0.9038208142
2	1	0	1	1	1	1		569.9229708	32.697129203
3	1	0	1	1	1	1		538.63763762	-25.23763762
4	1	0	1	1	1	1		570.87786484	-141.4178948
5	1	0	0	0	1	1		546.29092439	5.0590756148
6	1	0	0	1	1	1		568.12797625	-85.12797625
7	1	0	0	0	1	1		547.84509036	-1.685090359
8	1	0	0	1	1	1		553.66630233	160.37369767
9	1	0	1	1	1	1		577.0659721	-128.3959721
10	1	0	1	1	1	1		562.57589775	56.354112254
11	1	0	1	1	1	1		558.84071107	-57.38071107
12	1	0	0	0	1	1		558.62149589	99.208504108
13	1	0	1	1	1	1		585.6771188	5.6428811976

Figura 6-24

Los parámetros estarán almacenados en el fichero que para este propósito fue designado en el nodo de regresión y la pestaña *Output* (Figura 6-25).

VIEWTABLE: DMREG Parameter Estimates for node T28A8

	Row Type	Row Name	Intercept	actcult	aggrmad	deberes	decimo 0	edmad 1	edmad 2
1	PARMS	math	373.58779216	3.7670203309	26.560439687	-6.355332737	-54.37542435	-5.148187506	0.5627388
2	T		15.963080816	2.6170965275	10.281045636	-4.392227391	-15.63464269	-0.672276178	0.1261342175
3	LCLPARMS		327.71828701	0.9458694486	21.496995059	-9.191302318	-61.19194655	-20.15729007	-8.181500464
4	UCLPARMS		419.45729732	6.5881712133	31.623884306	-3.519363156	-47.55890315	9.8609150534	9.3069780643

Figura 6-25

EL NODO REGRESSION: MODELO LINEAL GENERAL GLM

Las diversas opciones de las pestañas de la pantalla *Linear and Logistic Regression* del nodo *Regression* (obtenida al hacer doble clic sobre el nodo *Regression*) que se observa en la Figura 6-26 permiten trabajar con el Modelo Lineal General GLM que, como ya sabemos, engloba todas las categorías de modelos de regresión incluidos los relativos al análisis de la varianza y la covarianza. Las pestañas de la Figura 6-26 permiten elegir los datos, las variables, las opciones, el método, la salida y otras características del modelo. La pestaña *Data* muestra las fuentes de datos disponibles que son predecesoras al nodo a la que podemos asignar uno de los papeles habituales:

- *Entrenamiento*: para ajustar el modelo.
- *Validación*: para valorar el modelo o para controlar el proceso de regresión por pasos sucesivos (*Stepwise*).
- *Test*: sirve para valorar el modelo.

©Alfaomega-RC Libros

- *Score*: sirve para predecir valores en una base de datos nueva que no contiene el objetivo. Para ello además de asignar a los datos el papel *Score*, en la subpestaña *Scored Data Set* de la pestaña *Output* activaremos el cuadrado *Score* (Figura 6-27).

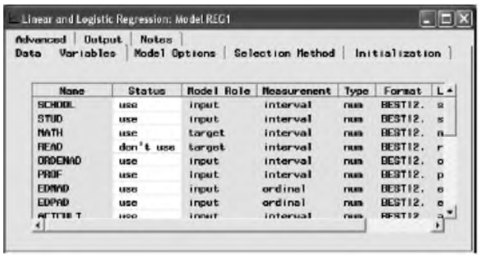


Figura 6-26

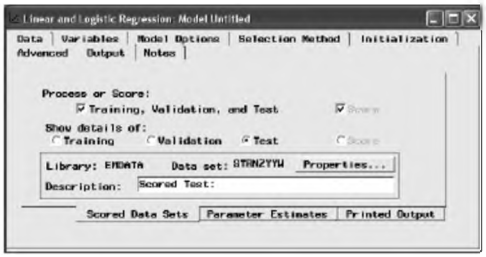


Figura 6-27

La pestaña *Variables* nos lleva a la tabla de la Figura 6-28, en la que podemos comprobar los parámetros como *Status, Model Role* o *Measurement* ya comentados en nodos anteriores. También podemos variar el estatus de una variable para que sea o no considerada en el análisis. Esto es de especial relevancia si tenemos varios *outputs* ya que solo podremos utilizar uno de ellos. Una de las funciones más interesantes de esta pestaña es la posibilidad de llamar a la función *Interaction Builder*. Cuando no existe un modelo teórico válido sobre todos los casos en el proceso de *Data Mining* es interesante indagar la posible influencia de efectos cruzados de determinadas variables que sospechemos que puedan tener una relación en el resultado con el objeto de implementar modelos completos del análisis de la varianza y la covarianza. Para activar esta opción seleccionaremos el icono para construir interacciones, que es el tercero en la barra de herramientas del nodo . Ello activará las opciones de la Figura 6-29.

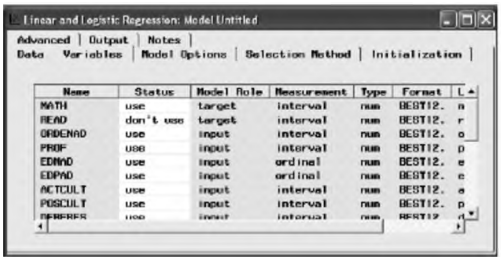


Figura 6-28

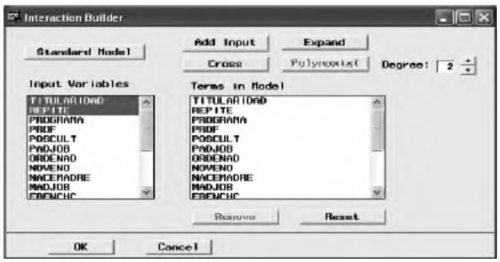


Figura 6-29

Standard Model: es el modelo por defecto que solo considera cada variable de forma individual. Para crear interacciones en la ventana *Input Variables* seleccionamos las variables que queramos combinar manteniendo pulsada la tecla *control*.

Cross: combina las variables seleccionadas.

***Expand**:* realiza todas las posibles combinaciones 2 a 2, 3 a 3, etc., de las variables seleccionadas.

Polynomial: crea los cuadrados, cubos, etc., de las variables seleccionadas.

Para borrar una variable una vez seleccionada haremos clic en *Remove*. Para volver al modelo sin interacciones haremos clic en *Reset*.

En la pestaña *Model Options* seleccionaremos las opciones de la Figura 6-30.

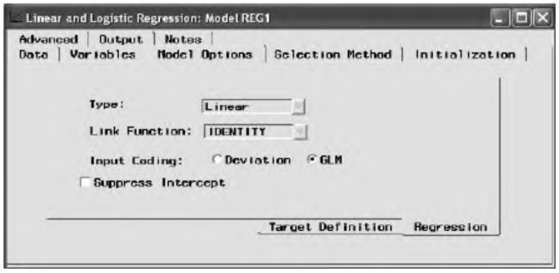


Figura 6-30

- *El tipo de regresión: lineal o logística.* Por defecto si la variable continua es dicotómica, se activará la opción logística, mientras que si es continua se activará la opción lineal.
- ***El método de codificar un input categórico:***
  - ***Deviation***: mediante esta codificación el parámetro estimado mide la diferencia entre cada nivel y la media a lo largo de cada nivel. Así, los parámetros para todos los niveles están restringidos a sumar cero. Esta codificación recibe también el nombre de *effects coding*. El siguiente ejemplo muestra cómo pueden ser codificadas las tres categorías de raza:

Código en desviaciones respecto a la media		
Level	Raza Blanca	Raza Negra
Blanco	1	0
Negro	0	1
Hispano	-1	-1

El parámetro estimado para Blanco y para Negro mide la diferencia del efecto entre ser de esa raza y la media de todos los demás niveles (Blanco, Negro, Hispano). Como la suma de los parámetros de todos los niveles es igual a cero, la estimación de la diferencia de los efectos entre la categoría Hispano y la media de todos los niveles calculando el valor negativo de la suma de los dos parámetros estimados.

©Alfaomega-RC Libros

- ***GLM:*** estamos ante el modelo Lineal General GLM, donde los parámetros estimados miden el efecto de la categoría con respecto a un nivel de referencia que por defecto es el último. Esta forma de codificación también recibe el nombre de variables dummy. El siguiente cuadro muestra cómo se realizaría una codificación de este tipo.

Código en deviciones respecto a la media		
Level	Raza Blanca	Raza Negra
Blanco	1	0
Negro	0	1
Hispano	0	0

- ***Supress Intercept***: permite eliminar la ordenada en el origen del modelo.

La pestaña ***Selection Method*** (Figura 6-31) permite construir el modelo mediante distintos métodos y criterios.

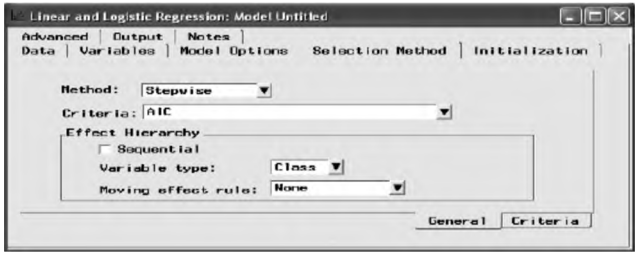


Figura 6-31

A continuación se describen los campos de la Figura 6-31 al seleccionar el botón *General* de la parte inferior derecha de la figura.

- ***Method***: existen cuatro opciones disponibles.
  - ***Backward***: parte de un modelo con todas las variables y efectos de los que irá eliminando aquellos no significativos hasta que ninguna variable cumpla el criterio de salida.
  - ***Forward***: parte de un modelo sin ninguna variable y de forma sistemática añade aquellos efectos más significativos hasta que ninguna variable cumpla el criterio de entrada.
  - ***Stepwise***: parte de un modelo sin ninguna variable y de forma sistemática añade y borra variables hasta que se cumplan los criterios de entrada y salida especificados.
  - ***None***: (por defecto): el analista define las variables que serán introducidas en el modelo.

La elección de un método por pasos sucesivos implica la selección de los criterios de entrada y salida de las variables. Para ello seleccionaremos la opción *Criteria* de la parte inferior derecha de la Figura 6-32.

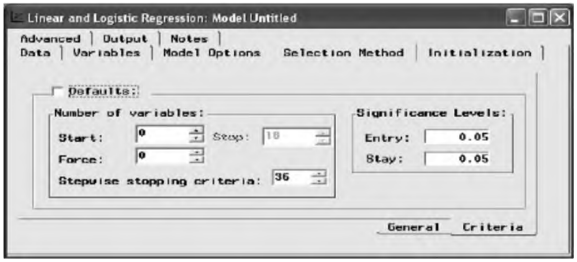


Figura 6-32

Por defecto, el nivel de significación se establece en el 95%. En la opción *Number of variables* se puede variar el número de variables para empezar el método de pasos sucesivos (*Start*), el número de parada (*Stop*) o forzar que exista un número mínimo de variables (*Force*). En la opción *Stepwise* se puede especificar un número de pasos máximos de selección de variables.

- ***Criteria:*** permite elegir el criterio para la selección del modelo. El nodo presenta las siguientes opciones.
  - ***AIC*** *(Akaike's Information Criterion):* el criterio de información de Akaike penaliza el número de parámetros del modelo.
  - ***SBC*** *(Schwarz's Bayesian Criterion):* el criterio bayesiano de Schwarz penaliza el número de parámetros del modelo. Se elegirá el modelo con el valor de SBC más reducido.
  - ***BIC*** *(Bayes' Information Criterion):* el criterio de información bayesiano. Penaliza el número de parámetros del modelo. Se elegirá el modelo con el valor de BIC más reducido.
  - ***Validation Error:*** elige el modelo con el error más bajo en la muestra de validación. Para los modelos de regresión logística el error es el valor negativo del *loglikelihood*. En regresiones múltiples el error es la suma de errores al cuadrado.
  - ***Validation Misclassification:*** elige el modelo con la tasa de error en la clasificación más reducida.
  - ***Cross-Validation Error:*** mediante esta medida se elige aquel modelo que tiene el error en la validación más bajo en la matriz de validación

©Alfaomega-RC Libros

cruzada de los datos de entrenamiento. En regresión logística el error es el negativo del loglikelihood. Para regresiones lineales el error es la suma de errores al cuadrado.

- ***Cross Validation Misclassification:*** elige el modelo con la tasa de error de clasificación más baja en la matriz de validación cruzada de los datos de entrenamiento.
- ***Profit/Loss:*** criterio por defecto si se definen dos o más decisiones en la matriz de pérdidas. El nodo elige el modelo que maximiza el error o minimiza la pérdida en los casos de la muestra de validación. Para usar este criterio es necesario definir una matriz de pérdida o de beneficios para el objetivo. Estos valores ajustan las probabilidades de pertenencia a priori que sean especificadas.
- ***Cross Validation Profit/Loss:*** este criterio elige el modelo que maximiza la matriz de beneficios o minimiza la matriz de pérdidas en la matriz de validación cruzada de los datos de entrenamiento.
- ***None:*** el modelo final se obtiene según el método de selección elegido.

- ***Effect Hierarchy:*** controla que si el modelo incluye valores cruzados entre dos variables también incluya las variables o efectos principales.
  - ***Sequential:*** fuerza la entrada secuencial de los efectos jerárquicos. Es decir, primero entrarían los efectos individuales, luego los cruzados y cuadrados y así sucesivamente. Así se previene que un modelo tenga efectos de grados elevados sin que al mismo tiempo estén los efectos individuales.
- ***Variable Type:*** determina si solo las variables de clase o también las variables de intervalo o continuas están sujetas a la opción de jerarquía.
- ***Moving Effect Rule:*** determina si la jerarquía se mantiene si todos los efectos individuales y múltiples salen y entran a la vez del modelo. Existen dos opciones:
  - ***Single:*** en esta opción solo puede entrar o salir del modelo un elemento al mismo tiempo de los sujetos a la jerarquía. En el primer paso se seleccionaría solo la variable A o la B. A continuación podría entrar el segundo efecto principal. Si ambos efectos se mantienen en el modelo como significativos, entonces podría entrar A*B. Antes de poder eliminar el efecto A o el B debería ser eliminado A*B.

- ***Multiple:*** en esta opción más de un efecto puede entrar o salir del modelo a la vez.
- ***None*** (por defecto): la jerarquía de los efectos no se mantiene y cualquier efecto puede entrar o salir en cualquier momento.

En la pestaña ***Initialization*** (Figura 6-33) se sitúan los valores de inicialización.

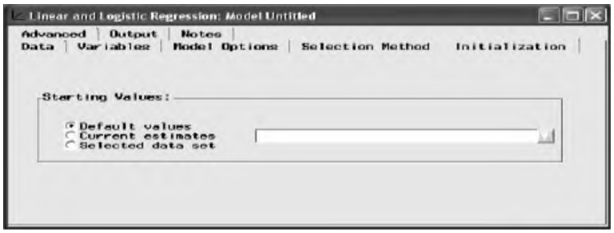


Figura 6-33

La pestaña de inicialización permite establecer los siguientes valores de comienzo (*Starting Values*):

Default values: los valores por defecto son asignar 0 tanto a las pendientes como a la ordenada en el origen. Estos son los valores para las estimaciones por máxima verosimilitud para el modelo con solo una constante.

Current estimates: se pueden usar los valores estimados de una regresión previa.

Selected data set: especifica un fichero que contiene los valores iniciales para estimar los parámetros. Este fichero debe ser compatible con los datos utilizados o el análisis no podrá ser ejecutado correctamente. Dado que el nodo no comprueba si los datos son o no compatibles es importante asegurarnos de que así sea.

La pestaña ***Advanced*** (Figura 6-34) permite indicar el algoritmo de optimización, el tiempo de uso máximo del microprocesador del ordenador y el criterio de convergencia para la regresión.

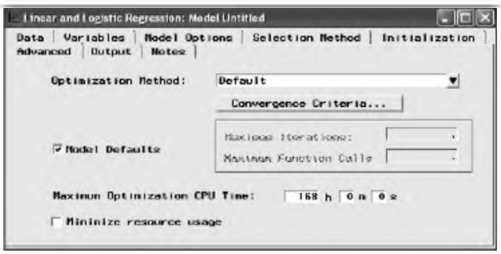


Figura 6-34

©Alfaomega-RC Libros

La siguiente tabla proporciona una descripción de los algoritmos de optimización no lineal (*Optimization Method*), número máximo de iteraciones permitidas (*Maximun Iterations*) y el número máximo de llamadas a la función (*Maximun Function Calls*).

Métodos de Optimización disponibles en el Nodo de Regresión.		
Algoritmo de Optimización	***Nº de Iteraciones Máximas***	***Nº Max. Llamadas a la Función***
Conjugate Gradient	400	1000
Double Dogleg	200	500
Newton-Raphson with Line Search	50	125
Newton-Raphson with Ridging	50	125
Quasi-Newton	200	500
Trust-Region	50	125

En principio el algoritmo debe ser elegido en función del tipo de problema de *Data Mining* que estemos llevando a cabo. La tipología es la siguiente:

- ***Problemas pequeños y medianos***: para problemas de hasta unos 40 parámetros, donde el hessiano es fácil de calcular, los mejores métodos son *Trust-Region, Newton-Raphson with line Search* y *Newton-Raphson with Ridging*.

- ***Problemas medianos***: para problemas donde el número de parámetros sea de unos 400, la función objetivo y el gradiente son mucho más fáciles de calcular que el hessiano. Es por ello que los métodos *Quasi-Newton* y *Double-Dogleg* requieren más iteraciones que los métodos anteriores pero su cálculo es mucho más rápido.

- ***Problemas grandes***: consideraremos que un problema con más de 400 parámetros es grande. En estos casos el algoritmo *Conjugate Gradient* es el más apropiado ya que este método es óptimo en problemas donde es necesario hacer un cálculo aproximado del hessiano.

El método por defecto depende del número de parámetros. Si es igual o menor a 40, el método elegido es *Newton-Raphson with Ridging*. Entre 41 y 400, el método por defecto es el *Quasi-Newton,* mientras que si el número de parámetros es mayor que 400 el nodo de regresión automáticamente selecciona el método *Conjugate Gradient*.

A pesar de que cada método presenta los números para la optimización vistos en la tabla anterior, estos pueden ser cambiados de la siguiente forma. En primer lugar es necesario deseleccionar el cuadro de diálogo *Model Defaults*.

A continuación se introducen los valores que consideremos oportunos. Por defecto el tiempo de optimización está establecido en 168 horas, valor que podemos cambiar en el campo especificado para esta tarea. En la regresión por pasos sucesivos el tiempo de optimización se reinicia cada vez que se ajusta un modelo.

Una lista de los criterios de convergencia (Figura 6-35) puede obtenerse haciendo clic en el botón *Convergence Criteria* de la Figura 6-34.

Estos valores por defecto pueden ser modificados introduciendo el nuevo valor en la celda correspondiente.

La pestaña *Output* nos lleva a una ventana en la que, además de las opciones habituales, disponemos de una subpestaña denominada *Printed Output* en la que podemos seleccionar outputs que por defecto no son mostrados (Figura 6-36).

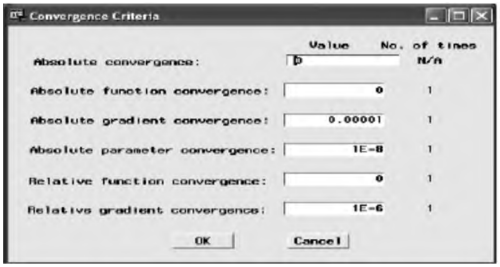


Figura 6-35

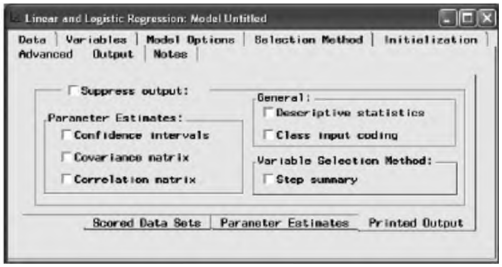


Figura 6-36

Una vez ejecutada la regresión obtendremos la *ventana de resultados* de la Figura 6-37 en la que está seleccionada la pestaña *Model,* que muestra todas las características del modelo ejecutado. En la pestaña *Estimates* podemos ver tanto gráfica como numéricamente el valor de los parámetros estimados. La ventana *Plot* permite construir distintos gráficos entre las variables y los resultados estimados.

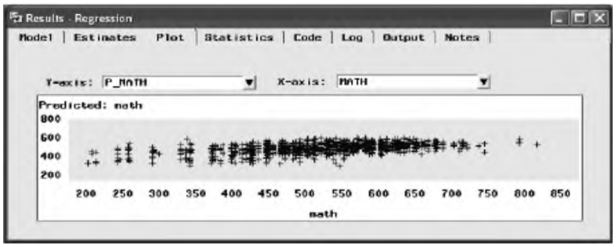


Figura 6-37

La ventana *Statistics* (Figura 6-38) muestra las estadísticas descriptivas de los criterios de parada y de optimización.

©Alfaomega-RC Libros

Model | Estimates | Plot | Statistics | Code | Log | Output | Notes

Fit Statistic	Label	Training	Validation	Test
AIC	Akaike's Information Criterion	12325.590377	.	.
ASE	Average Squared Error	6801.8565596	7454.0989298	7310.1017717
AVERR	Average Error Function	6801.8565596	7454.0989298	7310.1017717
DFE	Degrees of Freedom for Error	1393	.	.
DFM	Model Degrees of Freedom	3	.	.
DFT	Total Degrees of Freedom	1396	.	.
DIV	Divisor for ASE	1396	1047	1046
ERR	Error Function	9495112.5573	7804441.5795	7646366.4532
FPE	Final Prediction Error	6830.9523985	.	.
MAX	Maximum Absolute Error	315.76557312	301.69511627	331.264445
MSE	Mean Square Error	6816.3047791	7454.0989298	7310.1017717
NOBS	Sum of Frequencies	1396	1047	1046
_NW_	Number of Estimate Weights	3	.	.
RASE	Root Average Sum of Squares	82.472156269	86.337123706	85.499133163
RFPE	Root Final Prediction Error	82.649579642	.	.
RMSE	Root Mean Squared Error	82.560815566	86.337123706	85.499133163
SBC	Schwarz's Bayesian Criterion	12341.314475	.	.
SSE	Sum of Squared Errors	9495112.5573	7804441.5795	7646366.4532
SUMW	Sum of Case Weights Times Freq	1396	1047	1046

Figura 6-38

La ventana de *Output* (Figura 6-39) muestra los resultados obtenidos en el análisis.

Model | Estimates | Plot | Statistics | Code | Log | Output | Notes

Analysis of Variance

	DF	Sum of Squares	Mean Square	F Value	Pr > F
	2	3594201	1797101	263.65	<.0001
	1393	9495113	6816.304779	.	.
ed Total	1395	12089314	.	.	.

Model Fitting Information

R-square	0.2746	Adj R-sq	0.2735
AIC	12325.5904	BIC	12325.1403
SBC	12341.3145	C(p)	805.4919

The SAS System 10:38 Monday, February 28, 2005

The DMREG Procedure

Type III Analysis of Effects

ect	DF	Type III SS	F Value	Pr > F
cult	1	129851.6353	19.0502	<.0001
rmad	1	2919775.118	428.3516	<.0001

Analysis of Parameter Estimates

	DF	Estimate	Standard Error	t Value	Pr>|t|	95% Confidence Intervals Lower	Upper
	1	97.7312	25.2110	3.88	0.0001	48.3186	147.1
	1	10.5062	2.4300	4.36	<.0001	5.8435	15.3690
	1	70.1082	3.3874	20.70	<.0001	63.4690	76.7474

Figura 6-39

El nodo de regresión, además de los nodos *Ensemble* (combinación de modelos), *Tree* (árboles de decisión), *Neural Networks* (redes neuronales artificiales) y *User Defined Model* (modelo definido por el usuario), incluye una utilidad denominada *Model Manager*, en la cual se puede almacenar y acceder a distintos modelos a conveniencia. Esta utilidad permite obtener los mismos resultados que el nodo de valoración, por lo que trasladaremos el detalle de sus opciones a la discusión de este nodo (*Assessment node*). Para abrir el *Model Manager* existen dos opciones:

- Si el nodo está cerrado, hacer clic en el nodo de regresión y seleccionar *Model Manager*.

- Si el nodo está abierto, es necesario seleccionar en el menú principal *Tools* → *Model Manager*.

La pantalla de entrada del *Model Manager* presenta varias pestañas que se describen a continuación.

El *Model Manager* se abre en la pestaña de modelos *Models* (Figura 6-40), que muestra los modelos creados. De cada modelo se muestra información acerca de cuándo el modelo fue creado y qué estadísticas generó en los datos de entrenamiento, validación y test.

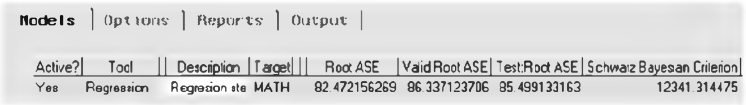


Figura 6-40

La pestaña *Options* sirve para seleccionar la partición de datos que usaremos para valorar el modelo, para especificar si los datos serán guardados después de la valoración y para indicar si haremos la valoración sobre el conjunto de datos o sobre una muestra (Figura 6-41).

La pestaña *Reports* muestra la valoración actual para el modelo seleccionado (Figura 6-42).

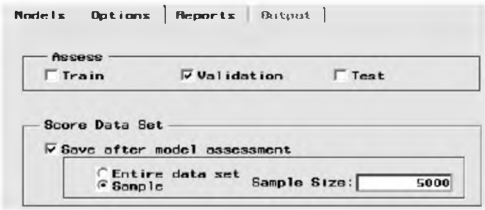


Figura 6-41

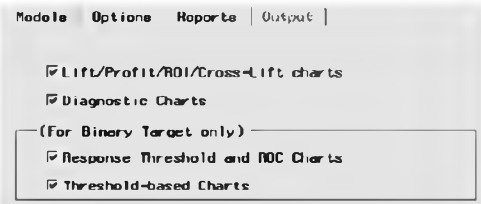


Figura 6-42

Para ver los gráficos (Figura 6-103) es necesario usar la opción *Tools* del menú principal y elegir una de las siguientes opciones dependiendo de los datos: *Lift Chart, Response, Threshold, ROC Chart*.

La pestaña *Output* solo aparece activa en el nodo de valoración para permitir pasar los resultados a otros nodos.

©Alfaomega-RC Libros

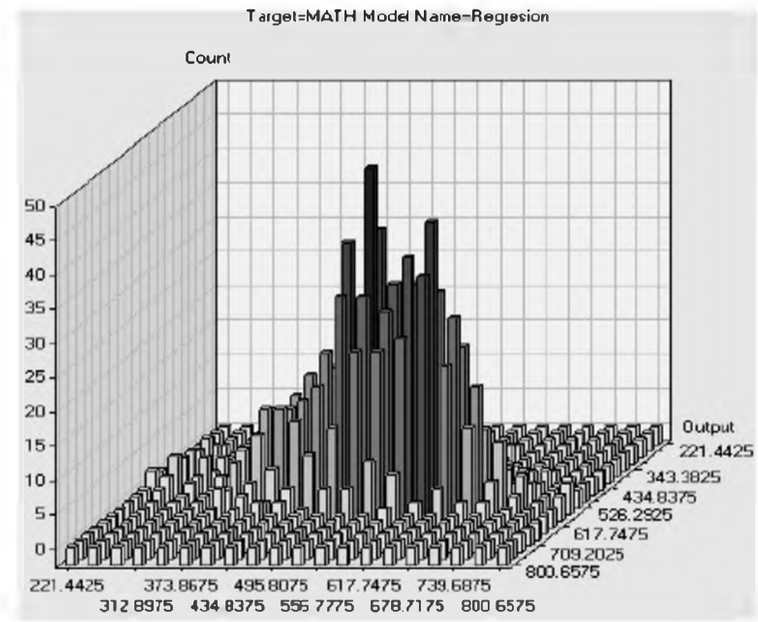


Figura 6-43

MODELOS DE ELECCIÓN DISCRETA LOGIT Y PROBIT CON EL NODO REGRESSION

A continuación llevaremos a cabo un ejercicio de regresión logística y la ejecución con el nodo de regresión. Para ello utilizaremos el fichero de datos *Creditos* donde analizaremos las variables explicativas que influyen en que un crédito sea o no pagado.

En primer lugar conectaremos el nodo que contiene los datos con el de regresión (Figura 6-44).

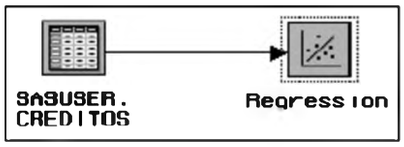


Figura 6-44

De forma análoga al ejemplo anterior, en la pestaña *Variables* activaremos las variables explicativas o inputs asignándoles un *Status* → *use* (Figura 6-45).

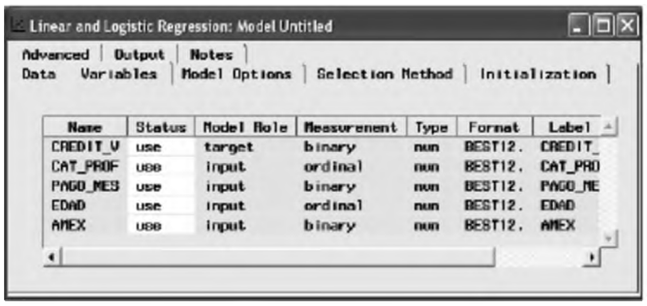


Figura 6-45

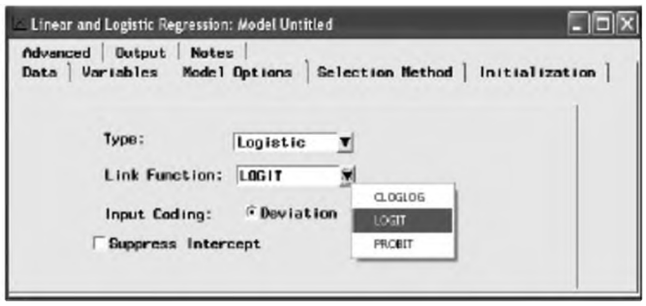


Figura 6-46

En la pestaña *Model Options* (Figura 6-46) seleccionaremos la opción *Logistic* y en este caso usaremos como función de transformación la función logística (LOGIT). Observamos que también puede elegirse la opción PROBIT.

Una vez seleccionadas las restantes opciones, como en el ejemplo anterior, (opciones por defecto) nombraremos el modelo y procederemos a su ejecución (Figura 6-47).

Figura 6-47

Los resultados obtenidos en el modelo de regresión logística pueden ser manipulados de forma análoga al caso de la regresión múltiple. Sin embargo, a diferencia del caso anterior, la variable dependiente es en esta ocasión dicotómica. Es por ello que la pestaña PLOT muestra ahora la matriz de confusión en forma de histogramas en tres dimensiones (Figura 6-48).

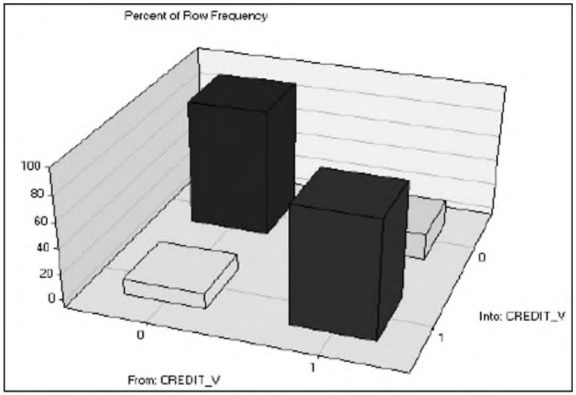


Figura 6-48

La pestaña *Output* muestra todo el detalle de resultados de la regresión logística. Una pequeña parte de estos es mostrada como ejemplo a continuación (Figura 6-49). Si queremos ver o analizar las predicciones de las probabilidades de pertenencia a grupo estimadas y los residuos obtenidos, estos están disponibles en el fichero creado guardado en la librería *EMDATA* y dentro de esta en el archivo *Strnqz*29 (Figura 6-50).

©Alfaomega-RC Libros

The SAS System 10:40 Monday, May 30, 20

The DMREG Procedure

Testing Global Null Hypothesis BETA=0

Criterion	Intercept Only	Intercept and Covariates	Chi-Square for Covariates
-2 LOG L	447.250	178.829	268.420 with 8 DF (p<.0001)

Type III Analysis of Effects

Effect	DF	Wald Chi-Square	Pr > Chi-Square
AMEX	1	0.6579	0.4173
CAT_PROF	4	8.1145	0.0875
EDAD	2	95.2061	<.0001
PAGO_MES	1	39.7557	<.0001

Analysis of Maximum Likelihood Estimates

Parameter		DF	Estimate	Standard Error	Wald Chi-square	Pr > Chi-square	exp(Est)
Intercept		1	1.7727	0.5654	9.83	0.0017	5.887
AMEX	0	1	0.1608	0.1982	0.66	0.4173	1.174
CAT_PROF	1	1	2.7304	1.0681	6.53	0.0106	15.340
CAT_PROF	2	1	-0.8562	0.5363	2.55	0.1104	0.425
CAT_PROF	3	1	-0.2442	0.5360	0.21	0.6486	0.783
CAT_PROF	4	1	-1.3729	0.8808	2.43	0.1188	0.253
EDAD	1	1	-2.4410	0.4365	31.27	<.0001	0.087
EDAD	2	1	0.0938	0.4789	0.00	0.9437	1.034
PAGO_MES	1	1	-1.6550	0.2625	39.76	<.0001	0.191

Figura 6-49

VIEWTABLE: Scored Train:

	Unnormalized Into: CREDIT_V	From: CREDIT_	Residual: CREDIT_V=1	Residual: CREDIT_V=0	Decision: CREDIT_V	Expected Profit: CREDIT_V	Best Profit: CREDIT_	Computed Profit: CREDIT_V	Predicted: CREDIT_V=1	Predicted: CREDIT_V=0
1	1	1	0.0799412311	-0.079841231	1	0.9201587689	1	1	0.9201587689	0.079841231
2	0	0	-0.3672842961	0.3672842961	1	0.3672842961	0	0	0.3672842961	0.632715704
3	0	0	-0.0206741102	0.0206741015	1	0.0206741015	0	0	0.0206741015	0.979325898
4	1	1	0.0591882714	-0.059188271	1	0.9408117286	1	1	0.9408117286	0.059188271
5	1	1	0.2884109181	-0.288410918	1	0.7115890819	1	1	0.7115890819	0.288410918
6	1	1	0.0277527973	-0.027752797	1	0.9722472027	1	1	0.9722472027	0.027752797
7	1	1	0.0058267483	-0.005826748	1	0.9941732517	1	1	0.9941732517	0.005826748
8	1	0	-0.5722857711	0.5722857113	1	0.5722857113	0	0	0.5722857113	0.427714289
9	0	0	-0.04658722	0.0465872196	1	0.0465872196	0	0	0.0465872196	0.95341278
10	0	0	-0.0826563866	0.0826563664	1	0.0826563664	0	0	0.0826563664	0.917343634
11	0	0	-0.0817576077	0.0817576073	1	0.0817576073	0	0	0.0817576073	0.918242393
12	0	0	-0.0206741102	0.0206741015	1	0.0206741015	0	0	0.0206741015	0.979325898
13	0	0	-0.4924179511	0.4924179506	1	0.4924179506	0	0	0.4924179506	0.507582049

Figura 6-50

CAPÍTULO 7

MODELADO PARA LA SEGMENTACIÓN Y PERFILADO CON HERRAMIENTAS DE IBM

TÉCNICAS DE MODELADO PARA LA SEGMENTACIÓN Y OBTENCIÓN DE PERFILES

La segmentación puede definirse como el proceso de dividir un todo (población, consumidores, etc.) en grupos uniformes más pequeños que tengan características semejantes denominados segmentos y que especifican un perfil. La segmentación suele realizarse según los valores de determinadas variables que son los que fijan sus características. Los segmentos son grupos homogéneos (por ejemplo, las personas en un segmento de una población son similares en sus actitudes sobre las variables de segmentación). Debido a esta similitud de los elementos dentro de cada grupo, es probable que respondan de modo similar a determinadas estrategias (marketing, ventas, precios, distribución, tratamientos, etc.). Entre los *requisitos* para una buena segmentación destacan la homogeneidad en el segmento, la heterogeneidad entre segmentos y la estabilidad de segmentos. Además, los segmentos deben ser identificables y medibles, deben ser accesibles y manejables y deben tener un tamaño adecuado para su rentabilidad.

La clasificación inicial de las técnicas de segmentación distingue entre técnicas predictivas, en las que las variables que intervienen en el proceso pueden clasificarse inicialmente en dependientes e independientes (similares a las técnicas del análisis de la dependencia o métodos explicativos del análisis multivariante) y técnicas descriptivas, en las que todas las variables tienen inicialmente el mismo estatus (similares a las técnicas del análisis de la interdependencia o métodos descriptivos del análisis multivariante).

Podríamos hacer una clasificación general de los tipos de segmentación en dos grandes categorías: segmentación Ad-Hoc y segmentación Post-Hoc.

En la segmentación Ad-Hoc, tanto el número de segmentos como su tamaño o su descripción se establece antes de que el estudio se lleve a cabo. En primer lugar, el investigador selecciona alguna base a partir de la cual segmentar (beneficio, necesidades, fidelidad de marca, etc.). En segundo lugar, el investigador clasifica a los individuos en segmentos de acuerdo con la base elegida y estudia su relación con otras variables descriptoras de tipo demográfico, socioeconómico, etc. La experiencia de los responsables y el conocimiento ayudan a la hora de conocer los segmentos importantes. Los árboles de decisión y el análisis discriminante son técnicas típicas de segmentación Ad-Hoc.

En la segmentación Post-Hoc no se conocen inicialmente el número de segmentos ni su tamaño. Esta situación se da, por ejemplo, cuando se desconocen las reacciones de los individuos ante los valores de las características de segmentación, en cuyo caso resulta más eficaz realizar una segmentación post hoc. En este modelo, el número de segmentos, su tamaño y su descripción se conocen tras el análisis y no antes.

En la segmentación Post-Hoc suele ser habitual realizar una exploración cualitativa para conocer en profundidad la población y, a continuación, aplicar un análisis de conglomerados que agrupa a los sujetos de acuerdo con la similitud de sus perfiles respecto a algunas variables de segmentación preestablecidas (comportamiento de compra, actitudes, beneficios buscados, etc). Esta segmentación también se denomina óptima, ya que permite determinar cuáles son los segmentos con la mayor homogeneidad interna y heterogeneidad entre ellos; es decir, determinar cuáles son los segmentos óptimos desde un punto de vista estadístico, con independencia de su mayor o menor ajuste a los requerimientos de la estrategia en que se inscriben.

©Alfaomega-RC Libros

A veces suele utilizarse un modelo híbrido entre los dos enfoques ad-hoc y post-hoc. El enfoque híbrido suele utilizarse en segmentaciones sucesivas basadas en primer lugar en características demográficas y después en variables tipo uso o razones de compra de un producto. El hándicap suele ser el elevado tamaño muestral necesario.

A continuación se presenta una *clasificación de las técnicas de segmentación*:

- Técnicas
  - Predictivas
    - Modelos de variable dependiente limitada
      - Elección discreta
      - Datos de recuento
    - Modelos censurados, truncados y de selección muestral
    - Modelos de duración, fallos y supervivencia
    - Modelos con ecuaciones estructurales
    - Análisis conjunto
    - Segmentación ad hoc
      - Modelo discriminante
      - Árboles de decisión
      - Redes neuronales
  - Descriptivas
    - Segmentación post hoc
      - Clustering
      - Redes neuronales
    - Reducción de la dimensión
      - Escalamiento Multidimensional
      - Análisis de correspondencias

EL ANÁLISIS CLUSTER COMO TÉCNICA DE PERFILADO Y SEGMENTACIÓN

El análisis de conglomerados, también llamado análisis cluster, constituye uno de los procedimientos estadísticos más utilizados hoy en día para la segmentación. La definición de la propia técnica se asemeja, de hecho, a los fines genéricos que persigue la segmentación: identificar grupos de sujetos lo más heterogéneos posible entre sí y lo más homogéneos posible dentro de cada grupo.

El análisis establece dichos grupos basándose en la similitud que presentan un conjunto de entidades (por ejemplo, turistas) respecto a una serie de características que el investigador ha de especificar previamente (motivaciones, necesidades, beneficios buscados, etc.). No obstante, es el análisis, y no el analista, el que finalmente extrae los grupos de sujetos y sus características definitorias: número de segmentos, número de integrantes de cada segmento, etc. El análisis de conglomerados constituye, por tanto, el ejemplo paradigmático del enfoque de *segmentación post hoc*. Asimismo se trata de un método descriptivo de segmentación.

El análisis cluster es una técnica de *Data Mining* de clasificación automática de datos. Su finalidad esencial es revelar concentraciones en los datos (casos o variables) para su agrupamiento eficiente en clusters (o conglomerados) según su homogeneidad. El agrupamiento puede realizarse tanto para casos como para variables, pudiendo utilizarse variables cualitativas o cuantitativas. Los grupos de casos o variables se realizan basándose en la proximidad o lejanía de unos con otras, por lo tanto es esencial el uso adecuado del concepto de distancia. Es fundamental que los elementos dentro de un cluster sean homogéneos y lo más diferentes posibles de los contenidos en otros clusters.

El análisis cluster es por tanto una técnica de clasificación, conociéndose también con el nombre de *taxonomía numérica*. Otros nombres asignados al mismo concepto son análisis de *conglomerados, análisis tipológico, clasificación automática* y otros. El número de clusters no es conocido de antemano y los grupos se crean en función de la naturaleza de los datos. Se trata por tanto de una técnica de clasificación *post hoc*. Podíamos definir el análisis cluster como un método estadístico multivariante de clasificación automática que a partir de una tabla de datos (casos-variables), trata de situarlos en grupos homogéneos, conglomerados o clusters, no conocidos de antemano pero sugeridos por la propia esencia de los datos, de manera que los individuos que puedan ser considerados similares sean asignados a un mismo cluster, mientras que individuos diferentes (disimilares) se localicen en clusters distintos.

La diferencia esencial con el análisis discriminante estriba en que en este último es necesario especificar previamente los grupos por un camino objetivo (técnica de clasificación *ad hoc*), ajeno a la medida de las variables en los casos de la muestra.

El análisis cluster define grupos tan distintos como sea posible en función de los propios datos sin especificación previa de los citados grupos (técnica de clasificación *post hoc*). Si las variables de aglomeración están en escalas muy diferentes será necesario estandarizar previamente las variables, o por lo menos trabajar con desviaciones respecto de la media. Es necesario observar también los valores atípicos y desaparecidos porque los métodos jerárquicos no tienen solución con valores perdidos y los valores atípicos deforman las distancias y producen clusters unitarios. También es nocivo para el análisis cluster la presencia de variables correlacionadas, de ahí la importancia del análisis previo de multicolinealidad. Si es necesario se realiza un análisis factorial previo y posteriormente se aglomeran las puntuaciones factoriales. La solución del análisis cluster no tiene por qué ser única, pero no deben encontrarse soluciones contradictorias por distintos métodos. El número de observaciones en cada cluster debe ser relevante, ya que en caso contrario puede haber valores atípicos que difuminen la construcción de los

 ©Alfaomega-RC Libros

clusters. Los conglomerados deben tener sentido conceptual y no cambiar mucho al variar la muestra o el método de aglomeración. Los grupos finales serán tan distintos como permitan los datos. Con estos grupos se podrán realizar otros análisis: descriptivos, discriminante, regresión logística, diferencia...

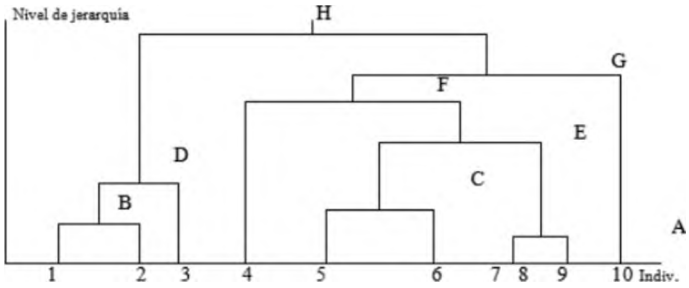
Clusters jerárquicos, secuenciales, aglomerativos y exclusivos (S.A.H.N.)

Los *métodos de análisis de conglomerados* que más se usan son los que son a la vez secuenciales, aglomerativos, jerárquicos y exclusivos, y que reciben el acrónimo, en lengua inglesa, de S.A.H.N. (*Sequential, Agglomerative, Hierarchic* y *Nonoverlaping*). En todos los *métodos de tipo S.A.H.N.* se siguen dos pasos fundamentales en el proceso de elaboración de los conglomerados. El primero de ellos es que los coeficientes de similitud *o* disimilitud entre los nuevos conglomerados establecidos y los candidatos potenciales a ser admitidos se recalcula en cada etapa, y el otro es el criterio de admisión de nuevos miembros a un conglomerado ya establecido. Entre los diferentes métodos de análisis de conglomerados de tipo S.A.H.N. tenemos los siguientes:

- Método de unión simple (*Single Linkage Clustering*), entorno o vecino más cercano (*Nearest Neighbour*) o método del mínimo (*Minimum Method*)
- Método de la distancia máxima o método del máximo (*Complete Linkage Clustering, Furthest Neighbour* o *Maximum Method*)
- Método de la media o de la distancia promedio no ponderado (*Weighted Pair Groups Method Using Arithmetic Averages* WPGMW)
- Método de la media ponderada o de la distancia Promedio Ponderado (*Group Average o Unweighted Pair Groups Method Using Arithmetic Averages* UPGMA)
- Método de la mediana o de la distancia mediana (*Weighted Pair Group Centroid Method* WPGMC)
- Método del Centroide o de la Distancia Prototipo (*Unweighted Pair Group Centroid Method* UPGMC)
- Método de Ward o de mínima varianza

El dendograma en el análisis cluster jerárquico

Es habitual en la investigación la necesidad de clasificar los datos en grupos con estructura arborescente de dependencia, de acuerdo con diferentes niveles de jerarquía. Partiendo de tantos grupos iniciales como individuos se estudian, se trata de conseguir agrupaciones sucesivas entre ellos de forma que progresivamente se vayan integrando en clusters los cuales, a su vez, se unirán entre sí en un nivel superior formando grupos mayores que más tarde se juntarán hasta llegar al cluster final que contiene todos los casos analizados. La representación gráfica de estas etapas de formación de grupos, a modo de árbol invertido, se denomina *dendograma* y se representa a continuación:



La figura, que corresponde a un estudio de los individuos, muestra cómo el 8 y el 9 se agrupan en un primer cluster (A). En un nivel inmediatamente superior, se unen los individuos 1 y 2 (cluster B); y enseguida los 5, 6 y 7 (C). Un paso siguiente engloba el cluster B con el individuo 3 (D); y así sucesivamente hasta que todos ellos quedan estructurados al conseguir, en el nivel más alto, el cluster total (H) que reúne los 10 casos.

Análisis cluster no jerárquico

La clasificación de todos los casos de una tabla de datos en grupos separados que configura el propio análisis proporciona clusters no jerárquicos. Esta denominación alude a la no existencia de una estructura vertical de dependencia entre los grupos formados y, por consiguiente, estos no se presentan en distintos niveles de jerarquía. El análisis precisa que el investigador fije de antemano el número de clusters en que quiere agrupar sus datos.

Como puede no existir un número definido de grupos o, si existe, generalmente no se conoce, la prueba debe ser repetida con diferente número a fin de tantear la

©Alfaomega-RC Libros

clasificación que mejor se ajuste al objetivo del problema, o la de más clara interpretación.

Los métodos no jerárquicos también se conocen como *métodos partitivos* o de optimización, dado que, como hemos visto, tienen por objetivo realizar una sola partición de los individuos en *K* grupos. Esto implica que el investigador debe especificar a priori los grupos que deben ser formados. Esta es, posiblemente, la principal diferencia respecto de los métodos jerárquicos. La asignación de individuos a los grupos se hace mediante algún proceso que optimice el criterio de selección. Otra diferencia está en que estos métodos trabajan con la matriz de datos original y no requieren su conversión en una matriz de proximidades. Pedret agrupa los métodos no jerárquicos en las cuatro familias siguientes*: reasignación, búsqueda de la densidad, directos y reducción de dimensiones*.

Los *métodos de reasignación* permiten que un individuo asignado a un grupo en un determinado paso del proceso sea reasignado a otro grupo en un paso posterior si esto optimiza el criterio de selección. El proceso termina cuando no quedan individuos cuya reasignación permita optimizar el resultado que se ha conseguido. Algunos de los algoritmos más conocidos dentro de estos métodos son el *método K-means (o K-medias)* de McQueen (1967), el *Quick Cluster Analysis* y el *método de Forgy*, los cuales se suelen agrupar bajo el nombre de *métodos centroides o centros de gravedad*. Por otra parte está el *método de las nubes dinámicas*, debido a Diday.

Los *métodos de búsqueda de la densidad* presentan una aproximación tipológica y una aproximación probabilística. En la primera aproximación, los grupos se forman buscando las zonas en las cuales se da una mayor concentración de individuos. Entre los algoritmos más conocidos dentro de estos métodos están el *análisis modal de Wishart*, el *método de Taxmap de Carmichael y Sneath*, y el *método de Fortin*. En la segunda aproximación, se parte del postulado de que las variables siguen una ley de probabilidad según la cual los parámetros varían de un grupo a otro. Se trata de encontrar los individuos que pertenecen a la misma distribución. Destaca en esta aproximación el *método de las combinaciones de Wolf*.

Los *métodos directos* permiten clasificar simultáneamente a los individuos y a las variables. Las entidades agrupadas, ya no son los individuos o las variables, sino que son las observaciones, es decir, los cruces que configuran la matriz de datos.

Los *métodos de reducción de dimensiones*, como el análisis factorial de tipo Q, guardan relación con el análisis cluster. Este método consiste en buscar factores en el espacio de los individuos, correspondiendo cada factor a un grupo. La

interpretación de los grupos puede ser compleja dado que cada individuo puede corresponder a varios factores diferentes.

Resulta muy intuitivo suponer que una clasificación correcta debe ser aquella en que la dispersión dentro de cada grupo formado sea la menor posible. Esta condición se denomina *criterio de varianza,* y lleva a seleccionar una configuración cuando la suma de las varianzas dentro de cada grupo (varianza residual) sea mínima.

Se han propuesto diversos algoritmos de clasificación no jerárquica, basados en minimizar progresivamente esta varianza, que difieren en la elección de los clusters provisionales que necesita el arranque del proceso y en el método de asignación de individuos a los grupos. Aquí se describen los dos más utilizados.

El *algoritmo de las H-medias* parte de una primera configuración arbitraria de grupos con su correspondiente media, eligiendo un primer individuo de arranque de cada grupo y asignando posteriormente cada caso al grupo cuya media es más cercana. Una vez que todos los casos han sido ubicados, calcula de nuevo las medias o centroides y las toma en lugar de los primeros individuos como una mejor aproximación de los mismos, repitiendo el proceso mientras la varianza residual vaya disminuyendo. La partición de arranque define el número de clusters que, lógicamente, puede disminuir si ningún caso es asignado a alguno de ellos.

El *algoritmo de las K-medias*, el más importante desde los puntos de vista conceptual y práctico, parte también de unas medias arbitrarias y, mediante pruebas sucesivas, contrasta el efecto que sobre la varianza residual tiene la asignación de cada uno de los casos a cada uno de los grupos. El valor mínimo de varianza determina una configuración de nuevos grupos con sus respectivas medias. Se asignan otra vez todos los casos a estos nuevos centroides en un proceso que se repite hasta que ninguna transferencia puede ya disminuir la varianza residual; o se alcance otro criterio de parada: un número limitado de pasos de iteración o, simplemente, que la diferencia obtenida entre los centroides de dos pasos consecutivos sea menor que un valor prefijado. El procedimiento configura los grupos maximizando, a su vez, la distancia entre sus centros de gravedad. Como la varianza total es fija, minimizar la residual hace máxima la factorial o intergrupos. Y puesto que minimizar la varianza residual es equivalente a conseguir que sea mínima la suma de distancias al cuadrado desde los casos a la media del cluster al que van a ser asignados, es esta distancia euclídea al cuadrado la utilizada por el método.

Como se comprueban los casos secuencialmente para ver su influencia individual, el cálculo puede verse afectado por el orden de los mismos en la tabla;

 ©Alfaomega-RC Libros

pese a lo cual es el algoritmo que mejores resultados produce. Otras variantes propuestas a este método llevan a clasificaciones muy similares.

Como cualquier otro método de clasificación no jerárquica, proporciona una solución final única para el número de clusters elegido, a la que se llegará con menor número de iteraciones cuanto más cerca estén las "medias" de arranque de las que van a ser finalmente obtenidas. Los programas automáticos seleccionan generalmente estos primeros valores, tantos como grupos se pretenda formar, entre los puntos más separados de la nube. Los clusters no jerárquicos están indicados para grandes tablas de datos, y son también útiles para la detección de casos atípicos: Si se elige previamente un número elevado de grupos, superior al deseado, aquellos que contengan muy escaso número de individuos servirían para detectar casos extremos que podrían distorsionar la configuración.

Es aconsejable realizar el análisis definitivo sin ellos, ya con el número deseado de grupos para después, opcionalmente, asignar los atípicos al cluster adecuado que habrá sido formado sin su influencia distorsionante. Un problema importante que tiene el investigador para clasificar sus datos en grupos es, como se ha dicho, la elección de un número adecuado de clusters. Puesto que siempre será conveniente efectuar varios tanteos, la selección del más apropiado al fenómeno que se estudia ha de basarse en criterios tanto matemáticos como de interpretabilidad. Entre los primeros, se han definido numerosos indicadores de adecuación como el Criterio cúbico de clusters y la Pseudo F que se describen en el ejemplo de aplicación práctica. El uso inteligente de estos criterios, combinado con la interpretabilidad práctica de los grupos, constituye el arte de la decisión en la clasificación multivariante de datos.

Matemáticamente, un método de clasificación no jerarquizado consiste en formar un número prefijado *K* de clases homogéneas excluyentes, pero con máxima divergencia entre las clases. Las *K* clases o clusters forman una única partición (*clustering*) y no están organizadas jerárquicamente ni relacionadas entre sí. La clasificación no jerárquica o de reagrupamiento tiene una estructura matemática menos precisa que la clasificación jerárquica. El número de métodos existentes ha crecido excesivamente en los últimos años y algunos problemas derivados de su utilización todavía no han sido resueltos.

CLUSTERING MEDIANTE REDES NEURONALES

Las técnicas de *clustering* se ocupan de realizar una partición de un conjunto dado de datos en varios grupos o categorías. Tienen aplicaciones en muchos campos como la medicina, la biología, la zoología y, sobre todo, en las tecnologías de la información.

La mayoría de las técnicas que se han desarrollado se puede dividir en tres grandes grupos: jerárquicas, de minimización de costes o distorsiones y de teoría de grafos. Sin embargo, cuando se trata de grandes conjuntos de datos, solamente las técnicas de minimización de costes son las adecuadas, por razones de complejidad computacional.

Dado un conjunto de datos $\{x_i \in R^n,\ i = 1, 2, \ldots, p\}$ que corresponden a *n* valores de un vector aleatorio *X* cuya función de distribución F_x se desconoce, pero se sabe que es una mixtura de *m* funciones de distribución, el problema consiste en formar *m* grupos o categorías con los datos según una cierta función de distancia o medida de distorsión, que en nuestro caso va a ser el error cuadrático (la distancia euclídea al cuadrado). Cada grupo va a venir representado por su centroide, es decir, su vector media. Por lo tanto, se trata de determinar *m* centroides, $\omega_i \in R^n,\ i = 1, 2, \ldots, m$, de manera que la distorsión entre cada dato y su centroide sea mínima, es decir, minimizar:

$$J(\omega_1, \omega_2, \ldots, \omega_m) = \sum_{i=1}^{n} \sum_{x_j \in C_i} \left\| x_j - \omega_i \right\|^2$$

La solución de este problema conduce a una partición del conjunto de datos en *m* clases, categorías o regiones, C_i, $i = 1, 2, \ldots, m$, definidas de la siguiente forma:

$$x_j \in C_i \Leftrightarrow \left\| x_j - \omega_i \right\| < \left\| x_j - \omega_k \right\|, \forall k \neq i$$

Obsérvese que, para una partición dada, los vectores de representación ω_i, $i = 1, 2, \ldots, m$, también llamados prototipos o vectores de referencia, que hacen mínima la distorsión son los centroides del grupo, es decir:

$$\omega_i = \frac{\sum_{j \in C_i} x_j}{n_i}$$

donde n_i es el número de elementos del grupo C_i.

El número de particiones diferentes de un conjunto de *N* elementos en *K* grupos es:

$$S(N, K) = \frac{1}{K!} \sum_{i=1}^{k} (-1)^{k-1} \binom{k}{i}^N$$

Para resolver este problema se han propuesto muchos algoritmos, como, por ejemplo, el algoritmo de las K-Medias de McQueen (1967). Sin embargo, estos algoritmos nos conducen a soluciones parciales, es decir, a mínimos locales, que no tienen por qué ser mínimos globales (solución óptima). Así, la dificultad de este problema radica en que la función objetivo tiene muchos mínimos locales, que no son necesariamente globales (solución óptima).

©Alfaomega-RC Libros

Para llevar a cabo este tipo de agrupaciones vamos a utilizar las redes neuronales competitivas con aprendizaje no supervisado, que permiten formar grupos o categorías con los datos de entrada. Cada entrada se asignará a la clase que le especifica la unidad de salida activada. En este tipo de aprendizaje las neuronas compiten entre ellas para ver cuál es la que se activa; solo una de ellas se activa, por ello, se le suele llamar el aprendizaje donde la ganadora se lleva todo. Aquí la red, por sí sola (sin supervisión), realiza la agrupación de los datos en categorías y nos da los prototipos de cada categoría. Este nuevo paradigma de aprendizaje fue introducido por Rumelhart y Zipser (1985), bajo el nombre de *aprendizaje competitivo*. La red consta de una capa de *m* neuronas o unidades de proceso que están conectadas con las señales o patrones de entrada y dichas conexiones tienen asociados unos pesos sinápticos que van a representar a los patrones prototipo. Cada neurona solo tiene dos posibles valores de salida: 0 (no activada) o 1 (activada) y solo se activa una neurona para cada patrón de entrada, aquella que recibe un mayor potencial sináptico (neurona ganadora).

Si la neurona i recibe la señal de entrada $x = (x_1, x_2, \ldots, x_n)'$ y tiene los pesos sinápticos $\omega_i = (\omega_{i1}, \omega_{i2}, \ldots, \omega_{in})'$, su potencial sináptico viene dado por la expresión:

$$h_i = \sum_{j=1}^{n} \omega_{ij} x_j - \frac{1}{2} \sum_{j=1}^{n} \omega_{ij}^2$$

Cuando $h_r = máx\{h_1, h_2, \ldots, h_m\}$ la salida de la red es:

$$(y_1, y_2, \ldots, y_{r-1}, y_{r+1}, \ldots, y_m) = (0,0,\ldots,0,1,0,\ldots,0)$$

Obsérvese que tal y como hemos definido el potencial sináptico se tiene que:

$$h_i < h_j \quad \Leftrightarrow \quad d(\omega_i, x) > d(\omega_j, x)$$

siendo d la distancia euclídea. Por lo tanto, la neurona que se activa es aquella cuyo vector de pesos sinápticos está más próximo al patrón de entra x, es decir, es el más parecido o similar.

Se trata de conseguir que el vector de pesos sinápticos de cada neurona llegue a ser el centroide (prototipo) del conjunto de datos que representa, es decir, para los cuales dicha neurona resulta ganadora. Ello se consigue con la *regla de aprendizaje competitivo*.

Para deducir la regla de aprendizaje competitivo vamos a considerar un conjunto de p patrones de entrada, que representaremos por $x_i = (x_{i1}, x_{i2}, \ldots, x_{in})'$, $i=1, 2, \ldots, p$, y al que llamaremos conjunto de entrenamiento de la red, pues a partir de él vamos a obtener el valor de los pesos sinápticos de cada neurona de la red. Si para la entrada x resulta ganadora la neurona es porque su vector de pesos es el más parecido y así esta entrada se asigna a la clase r.

Esto quiere decir que el vector de pesos sinápticos ω_r es el que representa al patrón de entrada *x*, y como deseamos que esta representación sea lo mejor posible, ω_r debe estar lo más próximo posible a todos los patrones del grupo que representa. Para ello debemos actualizar (modificar) el vector de pesos sinápticos ω_r de la neurona ganadora *r* acercándolo al patrón de entrada *s*. Es decir, si en la etapa *k*+1 hemos introducido el patrón de entrada *x* y ha resultado ganadora la neurona *r* entonces modificamos sus pesos sinápticos según la expresión:

$$\omega_r(k+1) = (1-\eta_r(k))\omega_r(k) + \eta_r(k)x(k) = \omega_r(k) + \eta_r(k)[x(k) - \omega_r(k)]$$

siendo η_r la tasa de aprendizaje de la neurona *r*.

Las demás neuronas no modifican sus pesos sinápticos. Esta es la regla estándar del aprendizaje competitivo. Obsérvese que es una combinación lineal entre el vector de entrada y el vector de pesos sinápticos y conforme menor sea el valor de la tasa de aprendizaje menor será la modificación de los pesos sinápticos. La tasa de aprendizaje de cada neurona debe de ser una función decreciente con respecto al número de la etapa, de forma que después de un número adecuado de etapas la red se estabiliza y deja de aprender. Durante el proceso de aprendizaje se introduce en cada iteración un patrón de entrada y se actualizan los pesos sinápticos, de manera que el proceso de aprendizaje finaliza después de introducir el conjunto de datos varias veces.

Con esta red, las neuronas aprenden individualmente a especializarse sobre conjuntos de patrones y son así detectoras de características. Cada vector de pesos sinápticos representa el prototipo de la clase o grupo que define la neurona correspondiente. Este algoritmo es similar al algoritmo de las K-Medias de McQueen, que consiste en ir asignando cada patrón al centroide más cercano y volver a determinar los centroides de los nuevos grupos formados. Ahalt y otros (1990) han puesto de manifiesto la conveniencia de utilizar el aprendizaje competitivo para resolver nuestro problema pues conduce a soluciones óptimas o cercanas a la óptima.

Ueda and Nakano (1994) han propuesto un nuevo aprendizaje competitivo con un mecanismo de selección basado en el principio de equidistorsión que permite a la red escapar de los mínimos locales y Uchiyama y Arbib (1994) han mostrado la relación existente entre nuestro problema de agrupación y la cuantificación vectorial, y presentan un algoritmo de aprendizaje competitivo que genera unidades donde la densidad de datos de entrada es alta, aplicándolo a la segmentación de imágenes en color.

 ©Alfaomega-RC Libros

TÉCNICAS DE SEGMENTACIÓN CON IBM SPSS MODELER

Dentro de la categoría *Segmentación* de la paleta *Modelado* (Figura 7-1), IBM SPSS Modeler dispone de los nodos *K-medias, Bietápico* y *Anomalía*, cuya finalidad es la realización de análisis *cluster* mediante el método K-medias, el método bietápico y el método de detección de anomalías. Asímismo se observa también el nodo *Kohonen* cuya finalidad es la segmentación clúster mediante redes neuronales.

Figura 7-1

El nodo K-medias: Cluster no jerárquico

El nodo *K-medias* de la categoría *Segmentación* de la paleta *Modelado* (Figura 7-1) permite realizar análisis *cluster* no jerárquico mediante el algoritmo de K-medias. Como ejemplo se consideran las variables *densidad, urbana, espvidaf, espvidam, alfabet, inc_pob, mortinf, tasa_nat, tasa_mort, tasasida* y *log_pib* del fichero *mundo.sas7bdat* y se trata de realizar un análisis *cluster* que agrupe los países del mundo según similitudes en estas variables que determinan el nivel de desarrollo.

Como el fichero de datos está en formato SAS, comenzamos utilizando el nodo *Archivo SAS* de la categoría *Orígenes* para situar como origen de datos el fichero *mundo.sas7bdat*.

Es una tarea previa al uso de cualquier modelo, definir el tipo de las variables que lo componen. El nodo *Tipo* permite asignar campos como variables dependientes o independientes en un modelo y no utilizar los que se desee. Para ello añadimos el nodo *Tipo* de la categoría *Operaciones con campos* al diagrama. A continuación hacemos clic con el botón derecho del ratón sobre el nodo *Tipo* y elegimos *Edición* en el menú emergente resultante. En la pantalla *Tipo* situamos como variables de entrada las variables *densidad, urbana, espvidaf, espvidam, alfabet, inc_pob, mortinf, tasa_nat, tasa_mort, tasasida* y *log_pib* de tipo *input* y las demás de tipo NINGUNO (Figura 7-2).

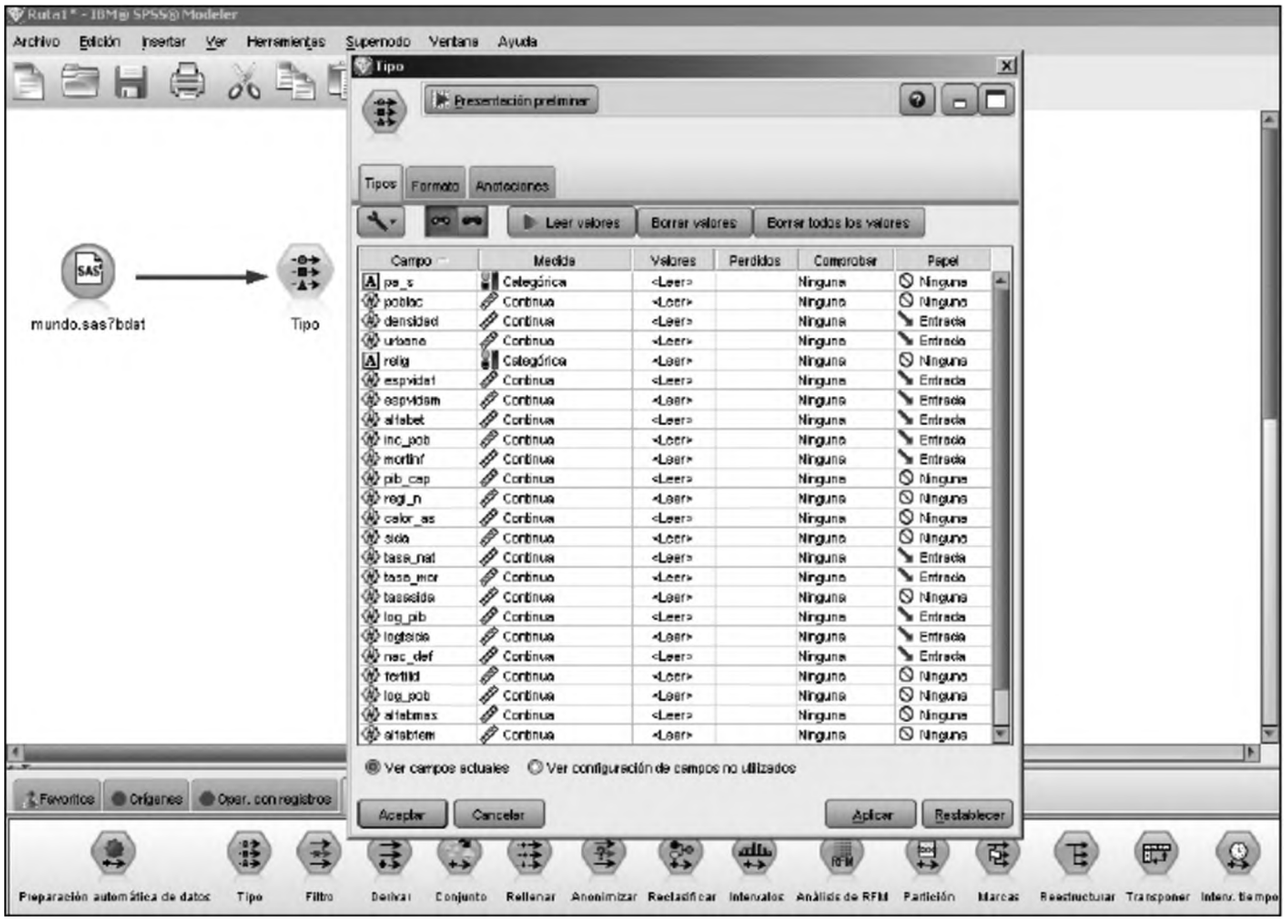


Figura 7-2

Para realizar el análisis de conglomerados añadimos el nodo *K-medias* de la categoría *Segmentación* de la ficha *Modelado* al diagrama. A continuación hacemos clic con el botón derecho del ratón sobre el nodo *K-medias* y elegimos *Edición* en el menú emergente resultante. Se obtiene la pantalla de la Figura 7-3 en cuya solapa *Campos* elegimos *Utilizar la configuración del nodo tipo*.

En la solapa *Modelo* elegimos *Automático, Utilizar los datos en particiones, Generar campo de distancia y Etiquetas de conglomerado Cadena* (Figura 7-4).

En la solapa *Experto* elegimos *Experto* y las opciones por defecto (Figura 7-5).

Al hacer clic en *Aplicar* y *Ejecutar* se obtiene el nodo K-medias sobre la ruta (Figura 7-6).

Si ahora hecemos clic con el botón derecho del ratón en el nodo K-medias sobre la ruta, se obtiene la pantalla de resultados de la segmentación K-medias de la Figura 7-7.

 ©Alfaomega-RC Libros

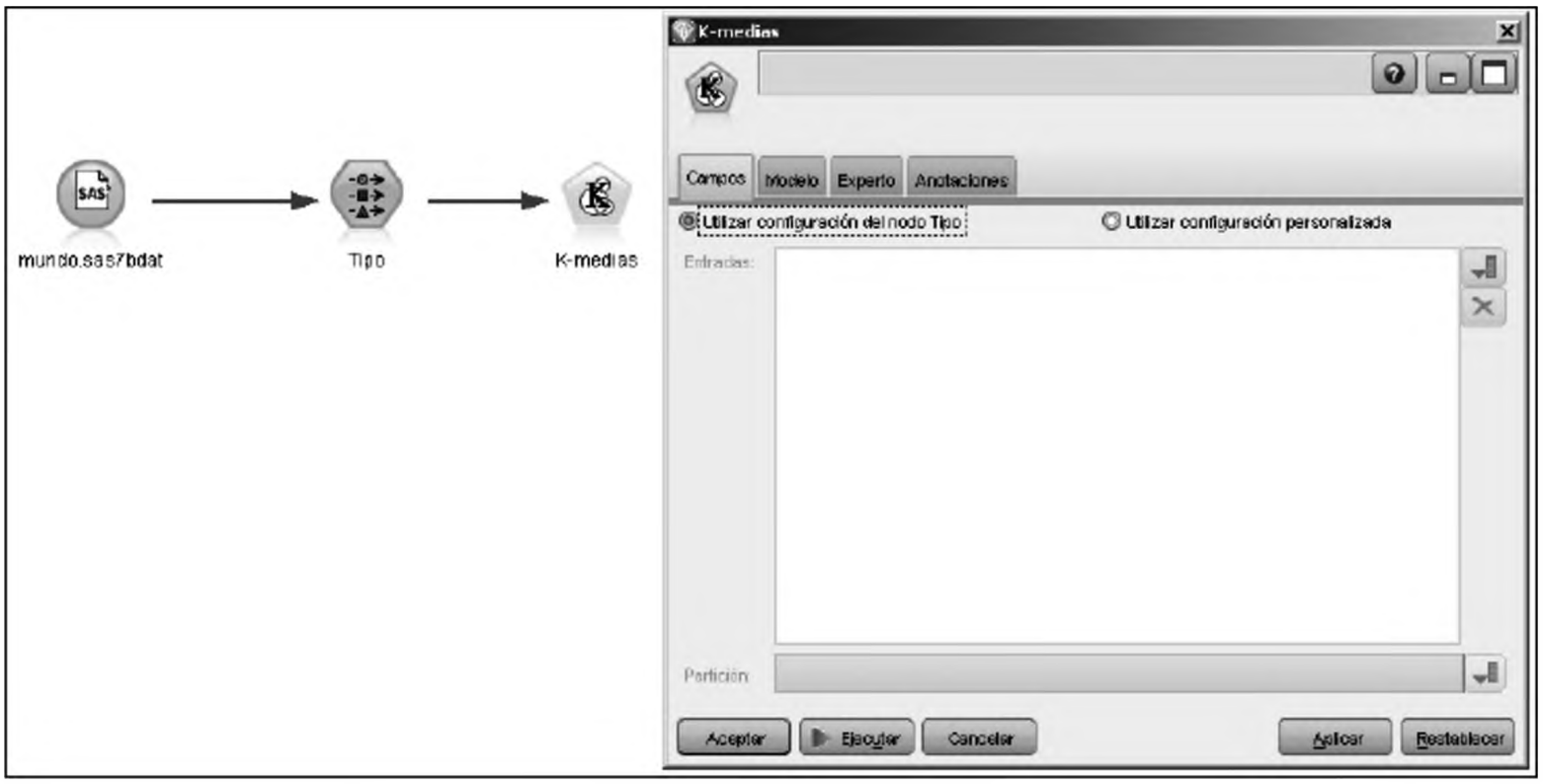


Figura 7-3

Figura 7-4

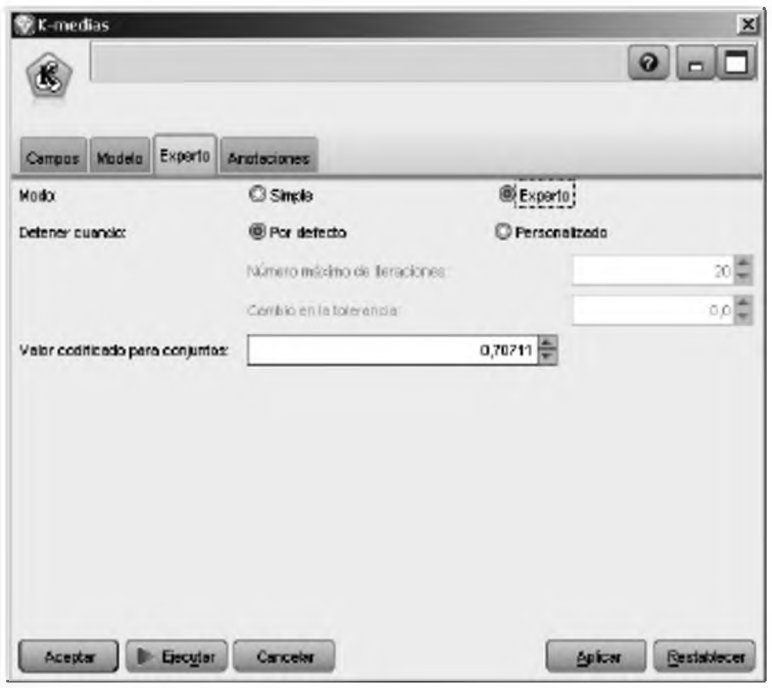


Figura 7-5

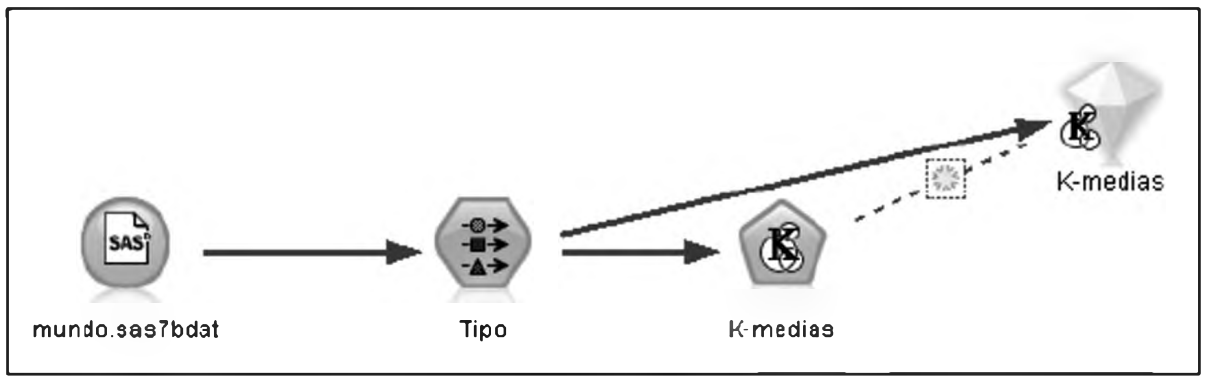


Figura 7-6

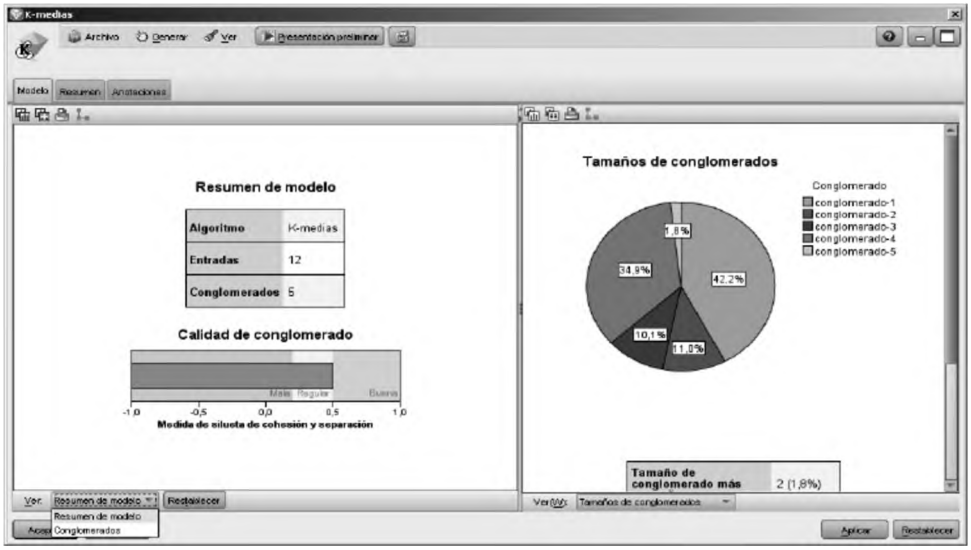


Figura 7-7

La vista *Resumen del modelo* muestra una instantánea o resumen del modelo de conglomerado, incluyendo una medida de silueta de la cohesión y separación de conglomerados sombreada para indicar resultados pobres, correctos o buenos. Esta instantánea le permite comprobar rápidamente si la calidad es insuficiente, en cuyo caso puede optar por volver al nodo de modelado para cambiar los ajustes del modelo de conglomerado para producir mejores resultados.

Los resultados serán pobres, correctos o buenos de acuerdo con el trabajo de Kaufman y Rousseeuw (1990) sobre la interpretación de estructuras de conglomerados. En la vista *Resumen del modelo*, un resultado "bueno" indica que los datos reflejan una evidencia razonable o sólida de que existe una estructura de conglomerados, de acuerdo con la valoración Kaufman y Rousseeuw; un resultado "correcto" indica que esa evidencia es débil, y un resultado "pobre" significa que, según esa valoración, no hay evidencias obvias.

Las medias de medida de silueta, en todos los registros, (B−A) / max(A,B), donde A es la distancia del registro al centro de su conglomerado y B es la distancia del registro al centro del conglomerado más cercano al que no pertenece. Un coeficiente de silueta de 1 podría implicar que todos los casos están ubicados directamente en los centros de sus conglomerados. Un valor de −1 significaría que todos los casos se encuentran en los centros de conglomerado de otro conglomerado. Un valor de 0 implica, de media, que los casos están equidistantes entre el centro de su propio conglomerado y el siguiente conglomerado más cercano.

 ©Alfaomega-RC Libros

El resumen incluye una tabla que contiene la siguiente información:

• Algoritmo. El algoritmo de conglomeración utilizado, por ejemplo, "Dos fases".

• Características de entrada. El número de campos, también conocidos como entradas o predictores.

• Conglomerados. Número de conglomerados de la solución.

La vista *Tamaños de conglomerados* (zona de la derecha de la Figura 7-7) muestra el gráfico de sectores que contiene cada conglomerado. El tamaño de porcentaje de cada conglomerado se muestra en cada sector, pase el ratón sobre cada sector para mostrar el recuento de ese sector.

Bajo el gráfico, una tabla enumera la siguiente información de tamaño:

• El tamaño del conglomerado más pequeño (un recuento y porcentaje del conjunto).

• El tamaño del conglomerado mayor (un recuento y porcentaje del conjunto).

• La proporción entre el tamaño del mayor conglomerado y el del menor.

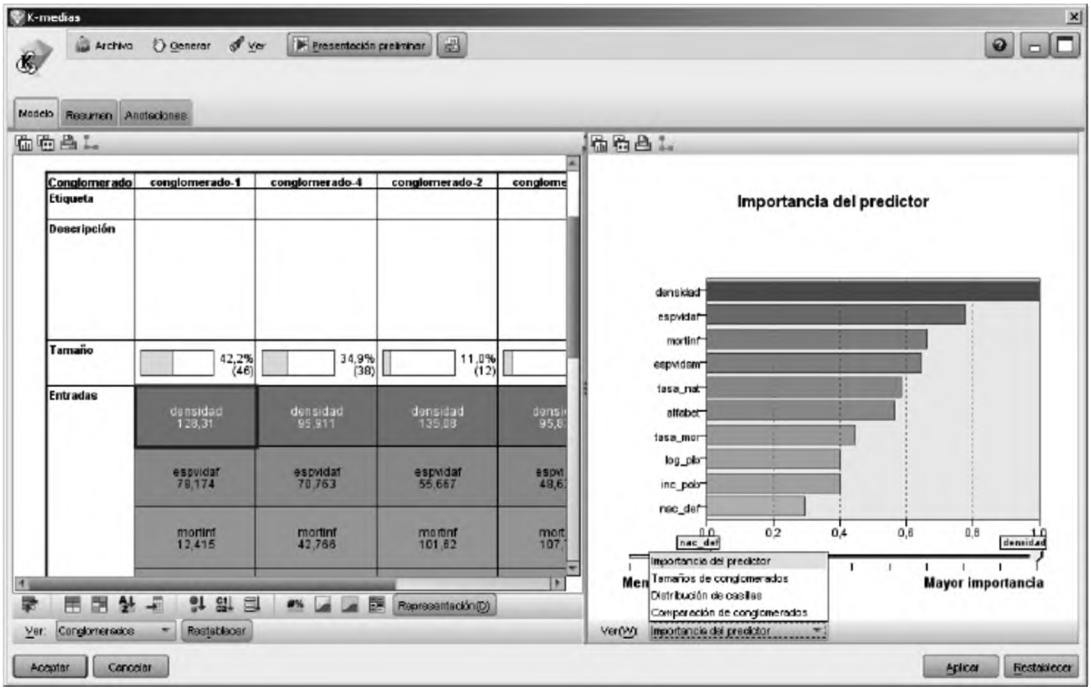


Figura 7-8

La vista *Conglomerados* (parte izquierda de la Figura 7-8) contiene una cuadrícula de conglomerados por funciones que incluye nombres de conglomerados, tamaños y perfiles para cada conglomerado.

Las columnas de la cuadrícula contienen la siguiente información:

• *Conglomerado*. Números de conglomerados creados por el algoritmo.

• *Etiqueta*. Etiquetas aplicadas a cada conglomerado (está en blanco de forma predeterminada). Pulse dos veces la casilla para introducir una etiqueta que describa el contenido del conglomerado, por ejemplo "Compradores de automóviles de lujo".

• *Descripción*. Cualquier descripción de los contenidos de los conglomerados (está en blanco de forma predeterminada). Pulse dos veces la casilla para introducir una descripción del conglomerado, por ejemplo "Más de 55 años de edad, profesionales, con ingresos superiores a 100.000 €".

• *Tamaño*. El tamaño de cada conglomerado como porcentaje de la muestra general del conglomerado. Cada casilla de tamaño de la cuadrícula muestra una barra vertical que muestra el porcentaje de tamaño del conglomerado, un porcentaje de tamaño en formato numérico y los recuentos de casos de conglomerado.

• *Funciones*. Los predictores o entradas individuales, ordenados por importancia general de forma predeterminada. Si hay columnas con tamaños iguales, se muestran en orden ascendente en función de los miembros del conglomerado.

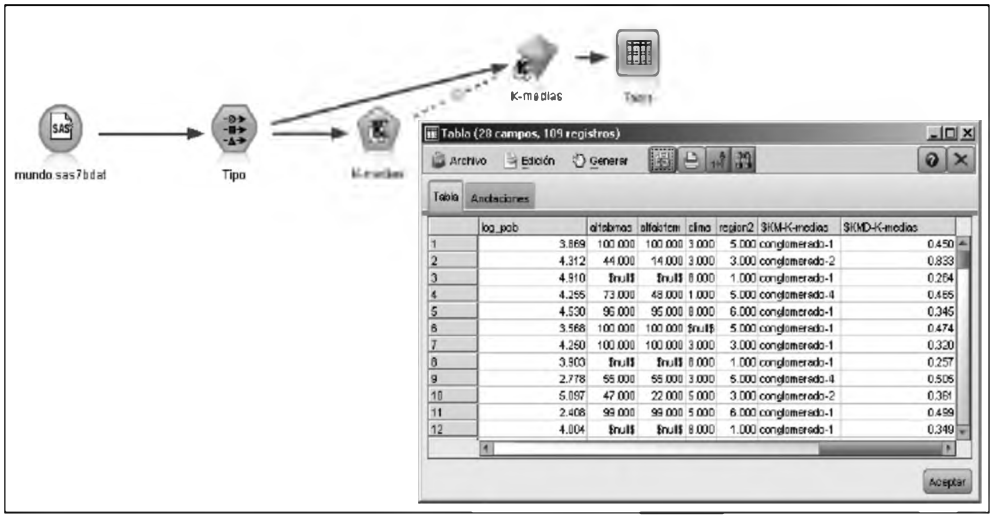
La importancia general de la característica se indica por el color del sombreado del fondo de la casilla, siendo más oscuro cuanto más importante sea la característica. Una guía sobre la tabla indica la importancia vinculada a cada color de casilla de característica.

Cuando pasa el ratón por una casilla, se muestra el nombre completo/etiqueta de la característica y el valor de importancia de la casilla. Es posible que aparezca más información, en función de la vista y tipo de característica. En la vista *Centros de conglomerados*, esto incluye la estadística de casilla y el valor de la casilla, por ejemplo: "Media: 4.32". En las características categóricas, la casilla muestra el nombre de la categoría (modal) más frecuente y su porcentaje.

La vista *Importancia del predictor* (parte derecha de la Figura 7-8) muestra la importancia relativa de cada campo en la estimación del modelo.

 ©Alfaomega-RC Libros

Si se añade un nodo tabla al nodo K-medias (Figura 7-9) se observan dos nuevos campos que contienen la pertenencia de cada registro (país) al conglomerado correspondiente y la distancia a partir del centro del conglomerado asignado para ese registro. Del nombre del modelo se derivan los nuevos nombres de campos con el prefijo $KM- para la pertenencia del conglomerado y $KMD- para la distancia desde el centro del conglomerado. Por ejemplo, si el modelo se llama Kmedias, los nuevos campos se llamarían $KM-Kmedias y $KMD-Kmedias.



	log_pob	altabmas	altaktem	clima	region2	$KM-K-medias	$KMD-K-medias
1	3.869	100.000	100.000	3.000	5.000	conglomerado-1	0.450
2	4.312	44.000	14.000	3.000	3.000	conglomerado-2	0.833
3	4.810	$null$	$null$	8.000	1.000	conglomerado-1	0.264
4	4.255	73.000	48.000	1.000	5.000	conglomerado-4	0.466
5	4.530	95.000	95.000	8.000	6.000	conglomerado-1	0.345
6	3.568	100.000	100.000	$null$	5.000	conglomerado-1	0.474
7	4.250	100.000	100.000	3.000	3.000	conglomerado-1	0.320
8	3.903	$null$	$null$	8.000	1.000	conglomerado-1	0.257
9	2.778	55.000	55.000	3.000	5.000	conglomerado-4	0.505
10	5.097	47.000	22.000	5.000	3.000	conglomerado-2	0.361
11	2.408	99.000	99.000	5.000	6.000	conglomerado-1	0.499
12	4.004	$null$	$null$	8.000	1.000	conglomerado-1	0.349

Figura 7-9

El nodo Bietápico: Cluster jerárquico

El nodo *Bietápico* de la categoría *Segmentación* de la paleta *Modelado* (Figura 7-1) permite realizar análisis *cluster* jerárquico mediante el algoritmo de las dos fases o método bietápico. Como ejemplo se consideran las variables *densidad, urbana, espvidaf, espvidam, alfabet, inc_pob, mortinf, tasa_nat, tasa_mort, tasasida* y *log_pib* del fichero *mundo.sas7bdat* y se trata de realizar un análisis *cluster* que agrupen los países del mundo según similitudes en estas variables que determinan el nivel de desarrollo.

Como el fichero de datos está en formato SAS, comenzamos utilizando el nodo *Archivo SAS* de la categoría *Orígenes* para situar como origen de datos el fichero *mundo.sas7bdat*.

Es una tarea previa al uso de cualquier modelo, definir el tipo de las variables que lo componen. El nodo *Tipo* permite asignar campos como variables dependientes

o independientes en un modelo y no utilizar los que se desee. Para ello añadimos el nodo *Tipo* de la categoría *Operaciones con campos* al diagrama. A continuación hacemos clic con el botón derecho del ratón sobre el nodo *Tipo* y elegimos *Edicion* en el menú emergente resultante. En la pantalla *Tipo* situamos como variables de entrada las variables *densidad, urbana, espvidaf, espvidam, alfabet, inc_pob, mortinf, tasa_nat, tasa_mort, tasasida* y *log_pib* de tipo *input* y las demás de tipo NINGUNO (Figura 7-10).

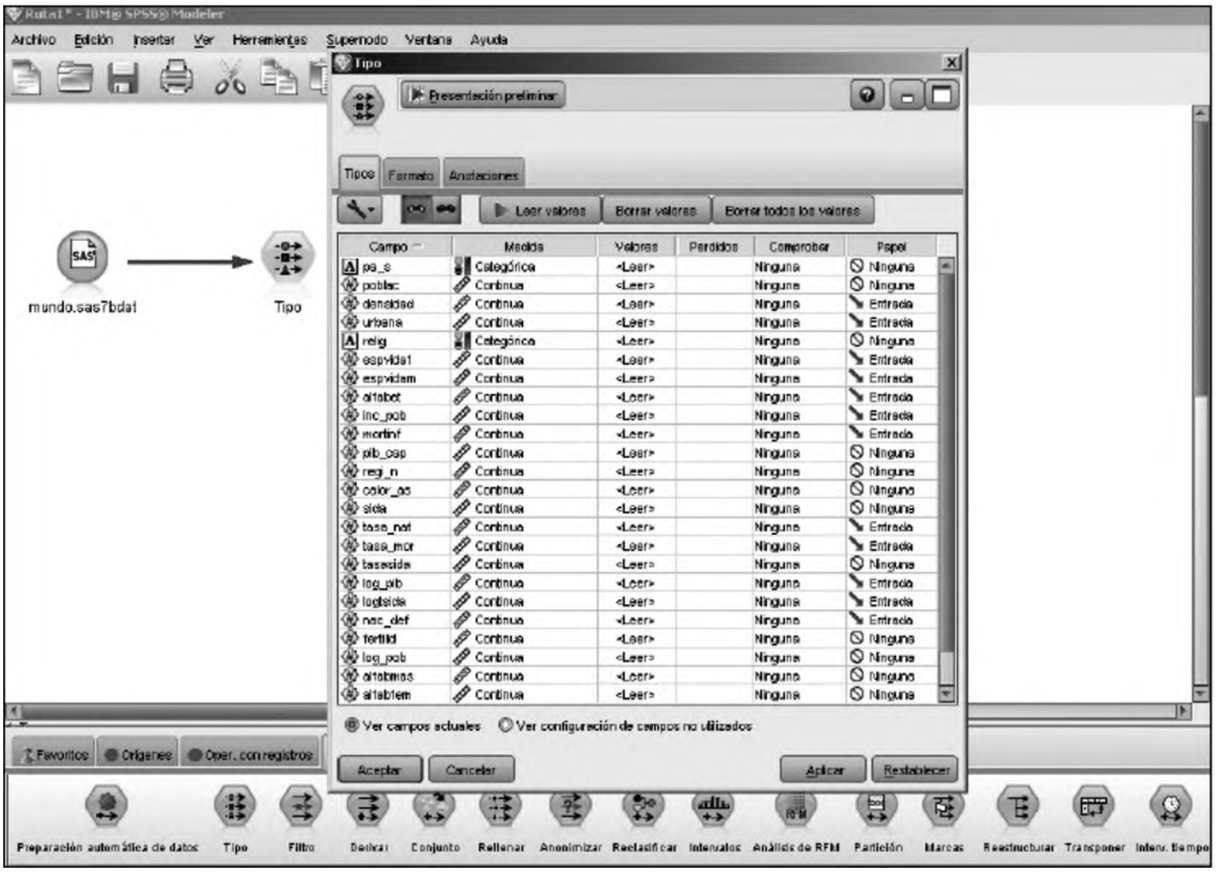


Figura 7-10

Para realizar el análisis de conglomerados añadimos el nodo *Bietápico* de la categoría *Segmentación* de la ficha *Modelado* al diagrama. A continuación hacemos clic con el botón derecho del ratón sobre el nodo *Bietápico* y elegimos *Edición* en el menú emergente resultante. Se obtiene la pantalla de la Figura 7-11 en cuya solapa *Campos* elegimos *Utilizar la configuración del nodo tipo*. En la solapa *Modelo* elegimos *Automático, Utilizar los datos en particiones, Estandarizar campos numéricos, Etiquetas de conglomerado Cadena, Calcular automáticamente número de conglomerados, Medición de la distancia* mediante *Log-verosimilitud y Criterio bayesiano de Swartz (BIC)* como *Criterio de conglomeración* (Figura 7-12). Al hacer clic en *Aplicar* y *Ejecutar* se obtiene el nodo Bietápico sobre la ruta (Figura 7-13).

©Alfaomega-RC Libros

Si ahora hecemos clic con el botón derecho del ratón en el nodo *Bietápico* sobre la ruta, se obtiene la pantalla de resultados de la segmentación bietápica de la Figura 7-14. Las distintas posibilidades de la salida (Figuras 7-14 y 7-15) se interpretan exactamente igual que en el caso del nodo K-medias visto anteriormente. Si se añade un nodo tabla al nodo Bietápico (Figura 7-16) se observan dos nuevos campos que contienen la pertenencia de cada registro (país) al conglomerado correspondiente y la distancia a partir del centro del conglomerado asignado para ese registro.

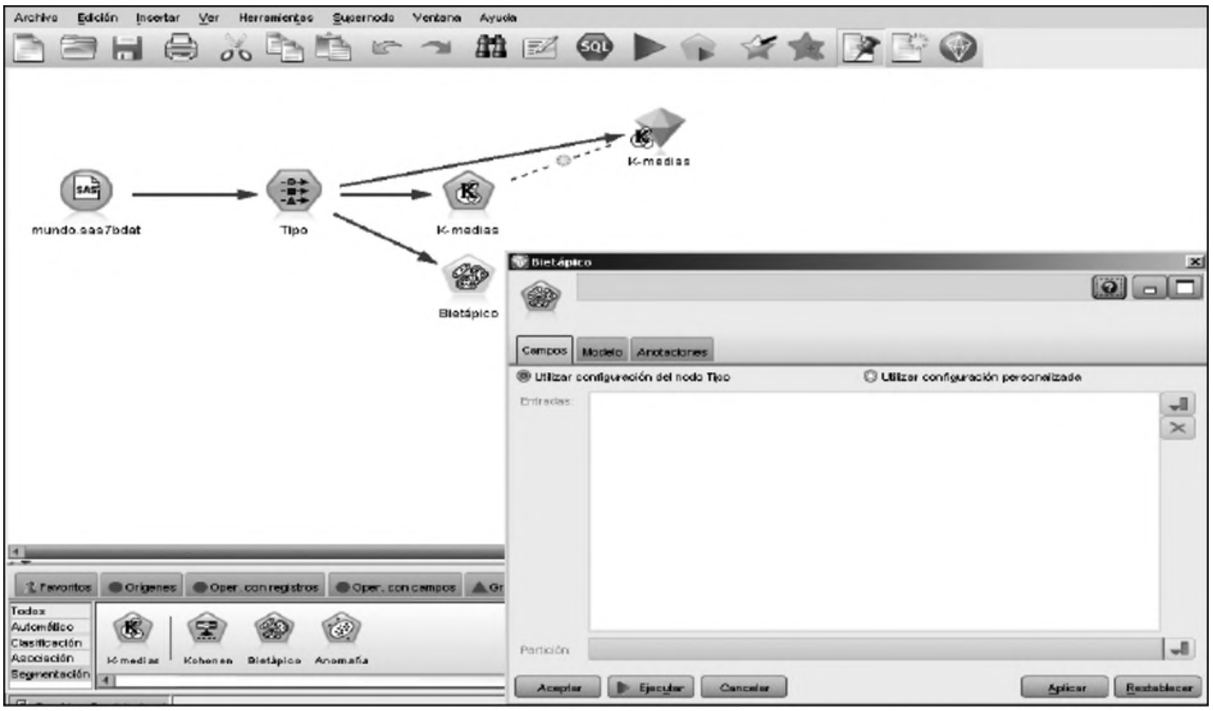


Figura 7-11

Figura 7-12

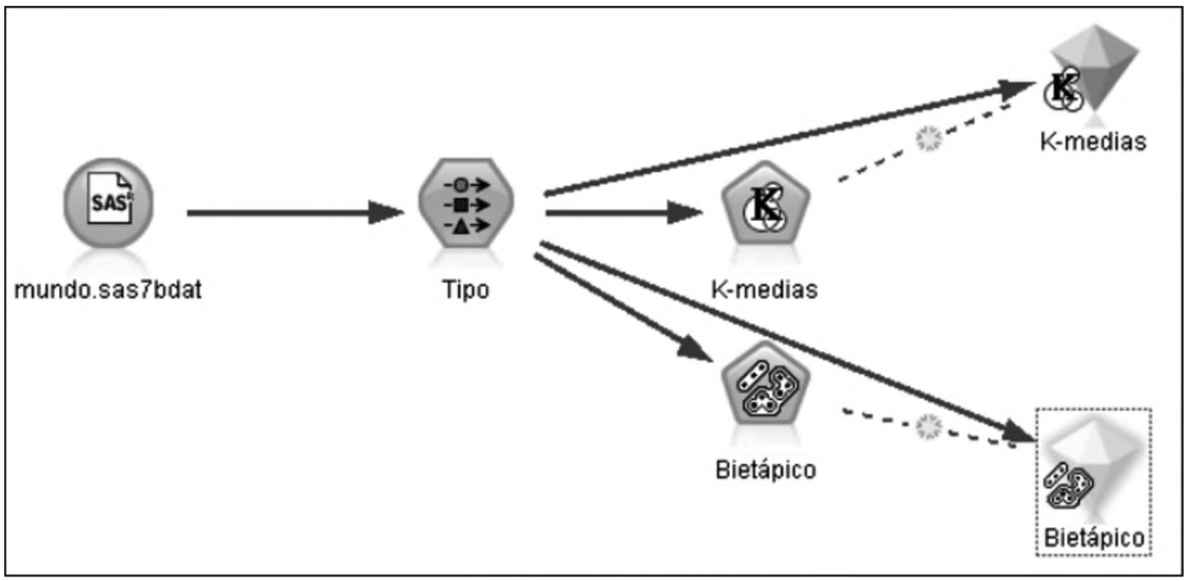


Figura 7-13

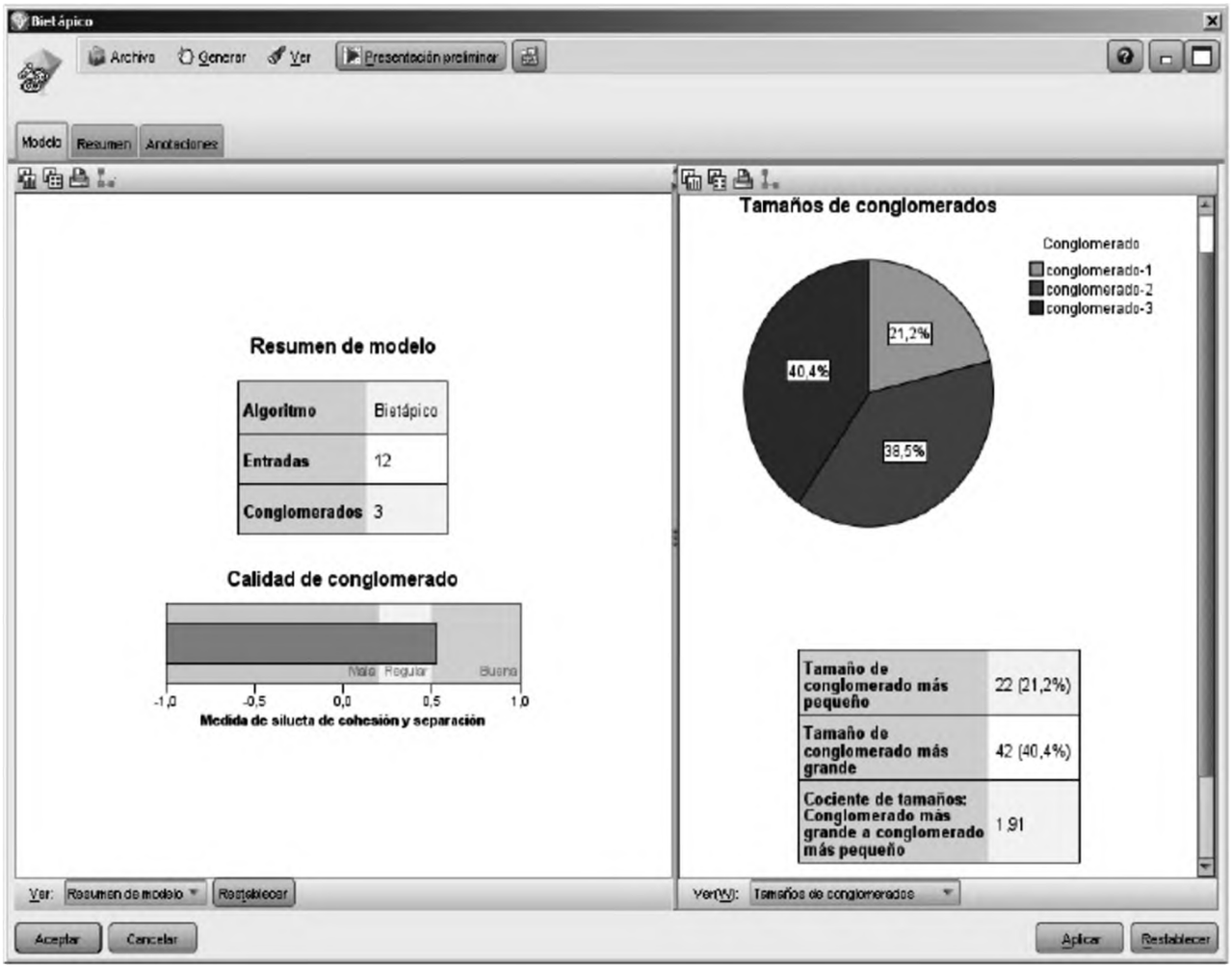


Figura 7-14

©Alfaomega-RC Libros

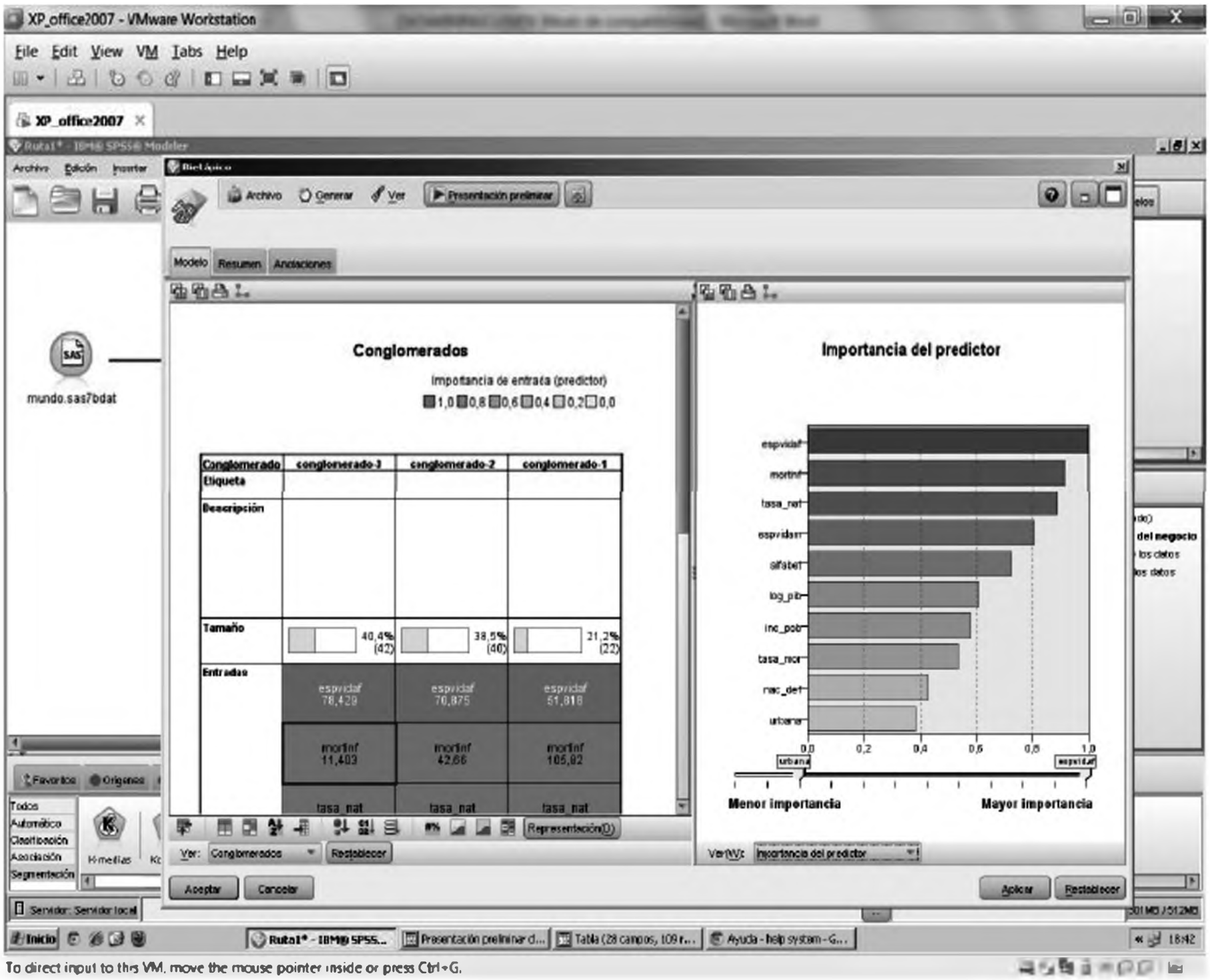


Figura 7-15

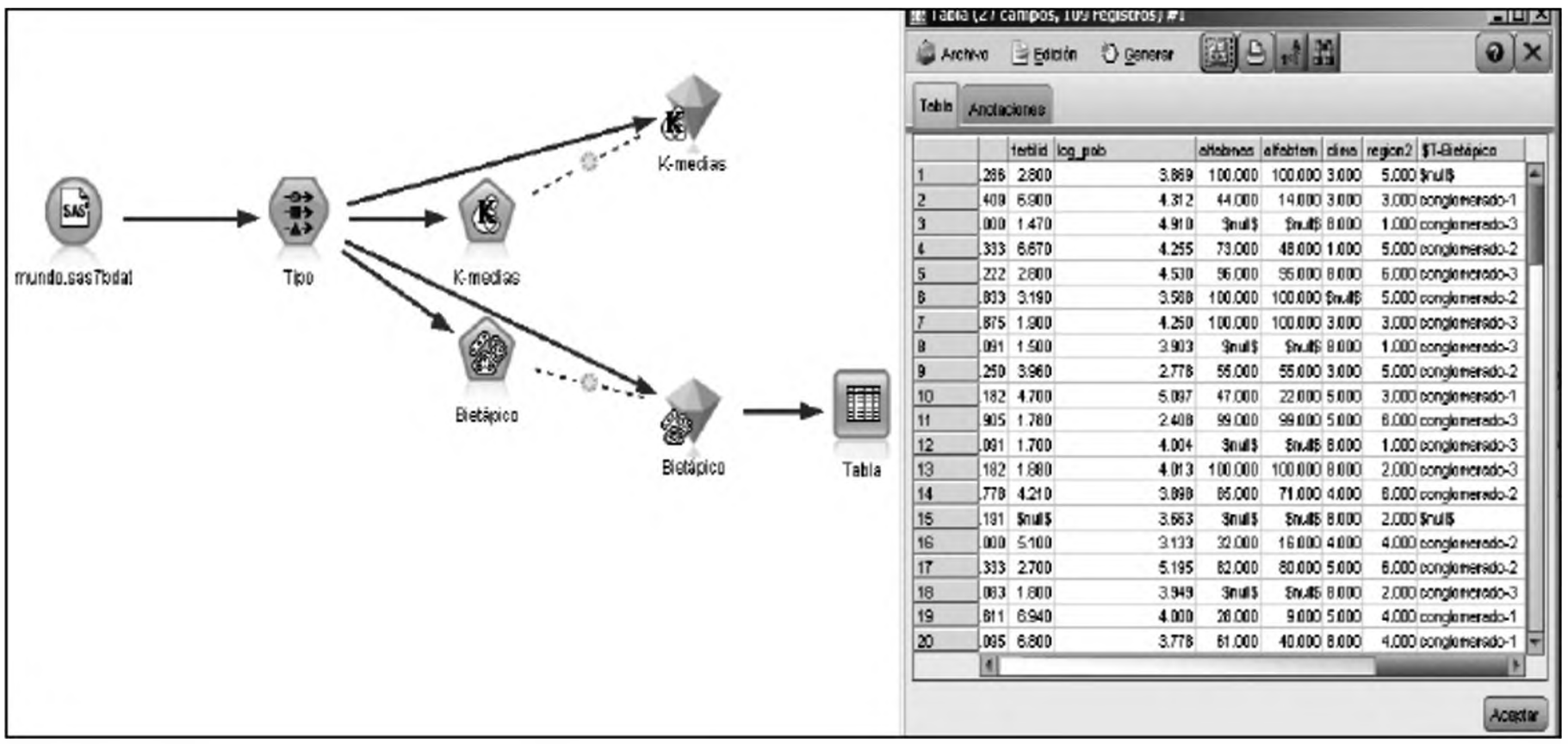


Figura 7-16

Análisis cluster con redes neuronales: Nodo Kohonen

El nodo *Kohonen* permite crear y entrenar redes neuronales de Kohonen, que suelen usarse para crear clusters cuando no se conoce el número inicial de grupos. Como ejemplo, vamos a utilizar los datos del archivo *Mundo* para agrupar, a partir de información de tipo social y económico, a distintos países del mundo en grupos homogéneos. El objetivo es conocer qué países son más o menos parecidos según las variables siguientes: *densidad, urbana, espvidaf, espvidam, alfabet, inc_pob, mortinf, tasa_nat, tasa_mort, tasasida* y *log_pib*. El resto de variables, salvo *PAÍS*, que tiene un papel de etiqueta y serán excluidas del análisis.

Comenzamos utilizando como origen de datos el conjunto de datos SAS *mundo.sas7bdat* y lo unimos con un nodo *Tipo* para situar como variables de entrada las involucradas en el análisis y como no utilizables el resto de las variables del archivo (Figura 7-17).

Para realizar el análisis de conglomerados añadimos el nodo *Kohonen* de la categoría *Segmentación* de la ficha *Modelado* al diagrama. A continuación hacemos clic con el botón derecho del ratón sobre el nodo *Kohonen* y elegimos *Edición* en el menú emergente resultante. Se obtiene la pantalla de la Figura 7-18 en cuya solapa *Campos* elegimos *Utilizar la configuración del nodo tipo*. En la solapa *Modelo* elegimos *Automático, Utilizar los datos en particiones, Mostrar gráfico de retroalimentación y Añadir etiqueta de conglomerado* (Figura 7-19). En la solapa *Experto* (Figura 7-20) elegimos *Experto* y el resto de las opciones por defecto. Al hacer clic en *Aplicar* y *Ejecutar* se obtiene el nodo *Kohonen* sobre la ruta (Figura 7-21).

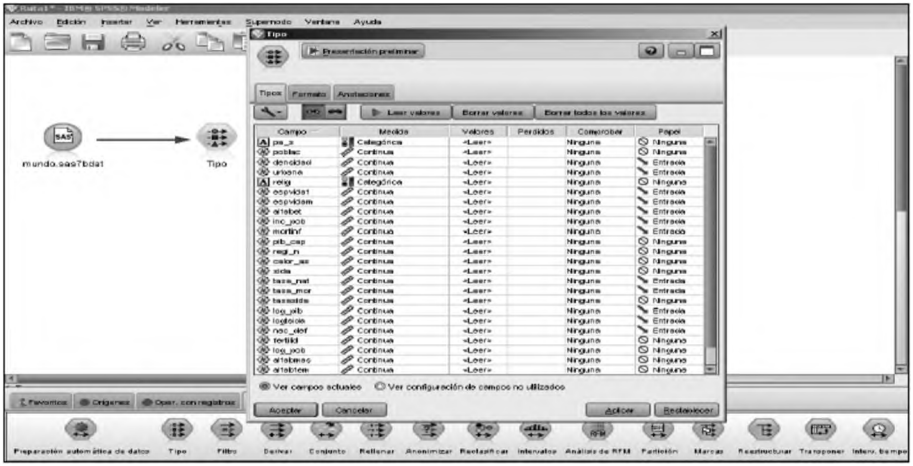


Figura 7-17

©Alfaomega-RC Libros

Figura 7-18

Figura 7-19

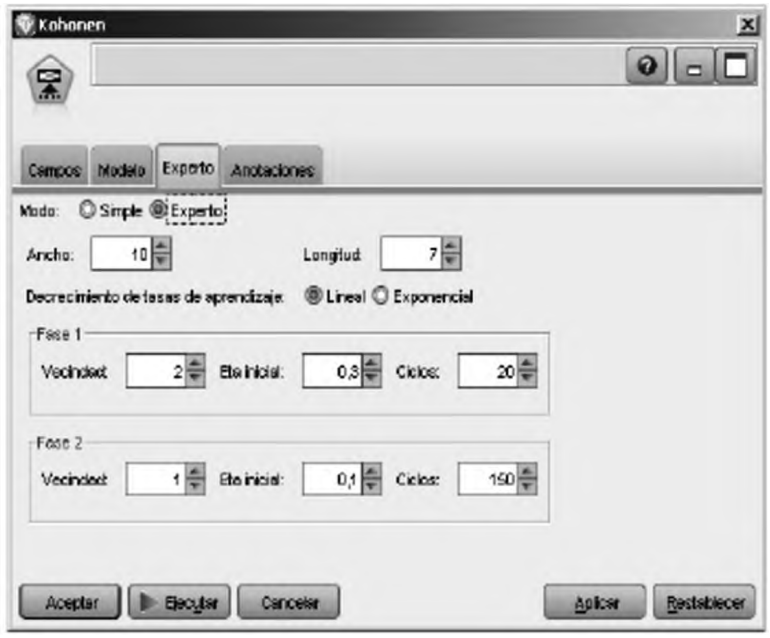


Figura 7-20

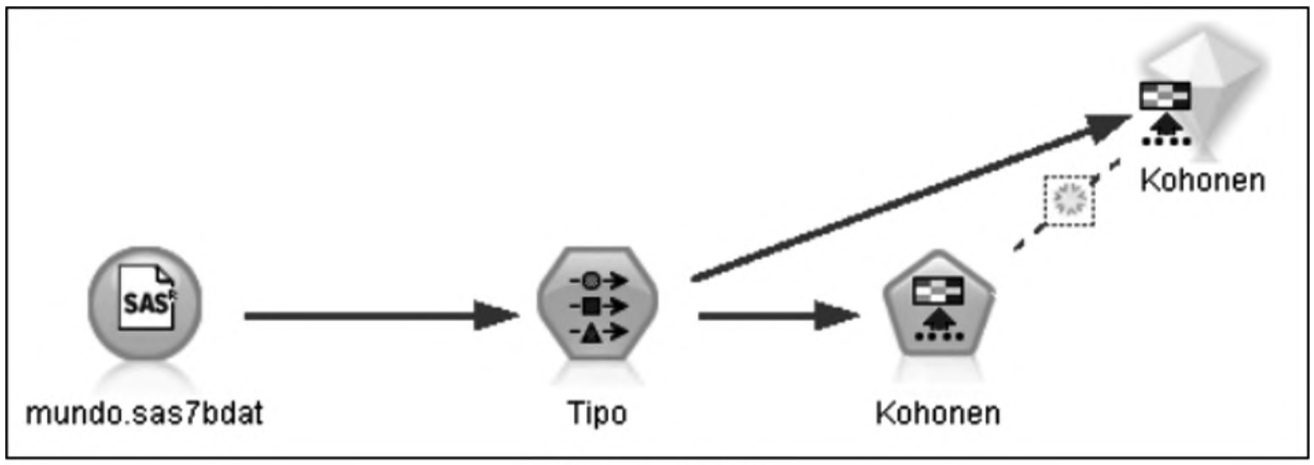


Figura 7-21

Si ahora hacemos clic con el botón derecho del ratón en el nodo *Kohonen* sobre la ruta, se obtiene la pantalla de resultados de la Figura 7-21. La pestaña *Modelo* de la Figura 7-21 muestra una visualización gráfica de estadísticas y distribuciones de resúmenes para campos entre conglomerados, que se conoce como el *Visor de conglomerados*.

El *Visor de conglomerados* se compone de dos paneles, la vista principal en la parte izquierda y la vista relacionada o auxiliar de la derecha. Hay dos vistas principales: *Resumen del modelo* (predeterminado) y *Conglomerados.*

Hay cuatro vistas relacionadas/auxiliares: *Importancia del predictor, Tamaños de conglomerados (predeterminado), Distribución de casillas* y *Comparación de conglomerados.*

En la Figura 7-21 se muestran las vistas *Resumen del modelo* y *Tamaño del conglomerado* y en la Figura 7-22 se muestran las vistas *Conglomerados* e *Importancia del predictor*. De igual forma se podría visualizar cualquier otro par de las posibles vistas.

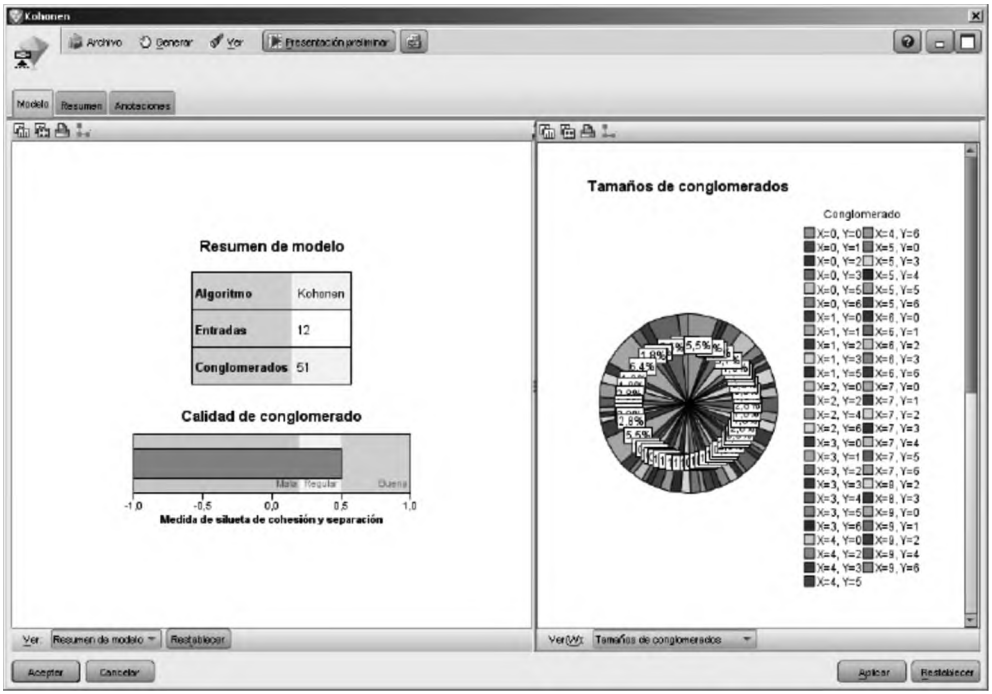


Figura 7-22

©Alfaomega-RC Libros

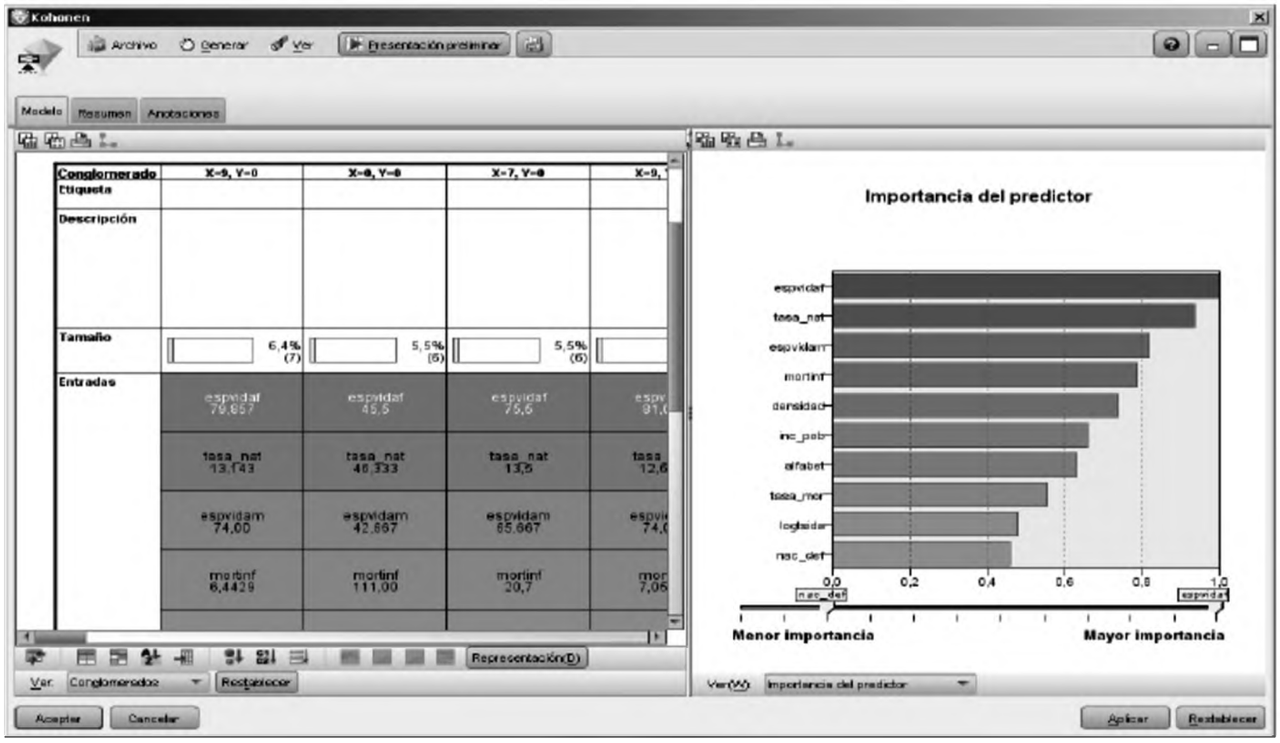


Figura 7-23

Si se añade un nodo tabla al nodo *Kohonen* (Figura 7-24) se observan dos nuevos campos que contienen las coordenadas X e Y de la unidad en la cuadrícula de resultados Kohonen que mejor respondieron a ese registro. Los nuevos nombres de campos se derivan del nombre del modelo, con los prefijos $KX- y $KY-. Por ejemplo, si el modelo se llama Kohonen, los nuevos campos se llamarían $KX-Kohonen y $KY-Kohonen. Si prefiere visualizar los conglomerados como una cuadrícula, puede consultar el resultado de la red de Kohonen trazando los campos $KX- y $KY- mediante un nodo Gráfico.

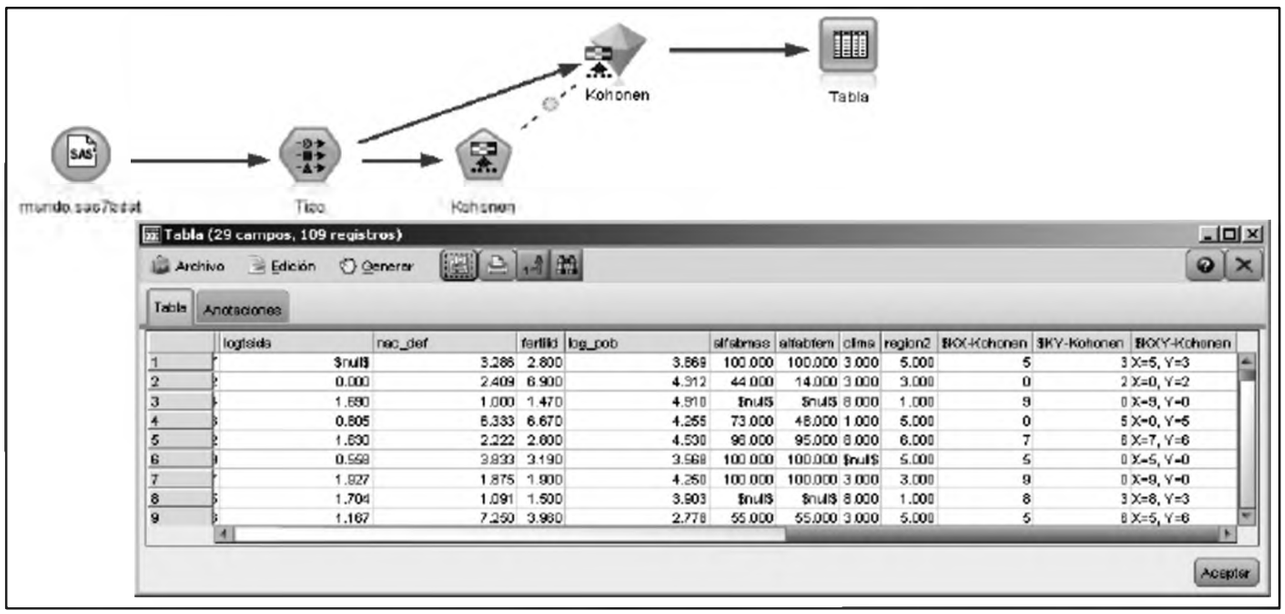


Figura 7-24

SEGMENTACIÓN Y PERFILADO CON ÁRBOLES DE DECISIÓN

Los árboles de decisión constituyen a la vez una herramienta predictiva y un método de segmentación, que trata de resolver los problemas de discriminación en una población segmentando de forma progresiva la muestra para obtener finalmente una clasificación fehaciente en grupos homogéneos, según la variable de interés denominada variable de segmentación.

Los árboles de decisión constituyen una técnica predictiva ya que la segmentación de la población se realiza según los valores de la variable de interés que juega el papel de variable dependiente del modelo predictivo subyacente en el árbol (variable cualitativa). La asignación de un elemento poblacional a un segmento se realiza de acuerdo a los valores de determinadas variables medidas sobre él que constituyen las variables independientes del modelo (habitualmente también variables cualitativas, aunque también suelen utilizarse variables cuantitativas con sus valores agrupados en un número pequeño de intervalos).

Se trata, por tanto, de seleccionar las variables explicativas que son más discriminantes para la variable dependiente y de construir una regla de decisión que permita asignar un nuevo individuo a un valor o clase de la variable dependiente. El método consiste en buscar la variable independiente x_j que mejor explique a la variable dependiente y. Esta variable define una primera división de la muestra en dos subconjuntos, llamados segmentos. Después se reitera el procedimiento en el interior de cada uno de estos dos segmentos buscando la segunda mejor variable y así sucesivamente.

Hay que observar que este método, contrariamente a otros métodos multidimensionales, no considera simultáneamente al conjunto de variables explicativas, sino que las examina una a una. Sin embargo, la relación entre variables explicativas se tiene en cuenta en las diferentes etapas del árbol.

De esta forma, construimos un árbol de decisión por divisiones sucesivas de la muestra en subconjuntos en el que se distinguen:

a) Los segmentos intermedios o nodos que engendran segmentos inmediatos.

b) Los segmentos terminales que no son divididos.

c) El árbol completo denotado A_{max} para el que cada segmento terminal contiene un solo individuo.

 ©Alfaomega-RC Libros

d) Un subárbol A que se obtiene a partir del anterior A_{max} por simple poda de una o muchas ramas.

Los árboles de decisión o árboles de clasificación, técnica muy utilizada hoy en día para segmentar, presentan de hecho un aspecto similar a los dendrogramas del análisis de conglomerados jerárquico, aunque se construyen e interpretan de forma completamente distinta. Los árboles de clasificación son en esencia particiones secuenciales del conjunto de datos realizadas para maximizar las diferencias de la variable dependiente o criterio base en los segmentos. Conllevan, por tanto, la división de las observaciones en grupos que difieren respecto a una variable de interés. Estos métodos se caracterizan además por desarrollar un proceso de división de forma arborescente. Mediante diversos índices y procedimientos estadísticos se determina la división más discriminante de entre los criterios seleccionados; es decir, aquella que permite diferenciar mejor a los distintos grupos del criterio base, obteniéndose de este modo la primera segmentación.

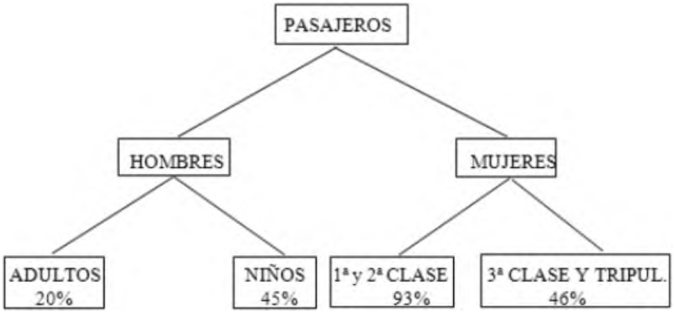
A continuación, se realizan nuevas segmentaciones de cada uno de los segmentos resultantes, y así sucesivamente hasta que el proceso finaliza con alguna norma estadística preestablecida o interrumpido voluntariamente en cualquier momento por el investigador. Además, los criterios descriptores no tienen por qué aparecer en el mismo orden para todos los segmentos, y un criterio puede aparecer más de una vez para un mismo segmento. Al final, enumerando los criterios mediante los que se ha llegado a un segmento determinado se obtiene el perfil del mismo.

Por ejemplo, supongamos que deseamos conocer qué pasajeros del Titanic tuvieron más probabilidades de sobrevivir a su hundimiento, y qué características estuvieron asociadas a la supervivencia al naufragio. En este caso, la variable de interés o variable dependiente es el grado de supervivencia. Podríamos entonces dividir a los pasajeros en grupos de edad, sexo y clase en la que viajaban y observar la proporción de supervivientes de cada grupo.

Un procedimiento arborescente selecciona automáticamente los grupos homogéneos con la mayor diferencia en proporción de supervivientes entre ellos; en este caso, el sexo (hombres y mujeres). El siguiente paso es subdividir cada uno de los grupos en función de otra característica, resultando que los hombres son divididos en adultos y niños, mientras que las mujeres se dividen en grupos basados en la clase en la que viajan en el barco.

Utilizar diferentes predictores en cada nivel del proceso de división supone una forma sencilla y elegante de manejar interacciones que a menudo complican en exceso los modelos lineales tradicionales. Cuando se ha completado el proceso de

subdivisión el resultado es un conjunto de reglas que pueden visualizarse fácilmente mediante un árbol. Por ejemplo: si un pasajero del Titanic es hombre y es adulto, entonces tiene una probabilidad de sobrevivir del 20 por ciento. Además, la proporción de supervivencia en cada una de las subdivisiones puede utilizarse con fines predictivos para vaticinar el grado de supervivencia de los miembros de ese grupo. Un árbol de clasificación del grado de supervivencia de los pasajeros del Titanic podría ser el que se observa en la figura siguiente.



CARACTERÍSTICAS DE LOS ÁRBOLES DE DECISIÓN

Podríamos enumerar como características más importantes en un árbol de decisión las siguientes:

Especificación de los criterios para minimizar los costes. Se trata de clasificar o predecir con el coste mínimo. Los costes suelen venir medidos en términos de la proporción de casos mal clasificados y uso inadecuado de probabilidades a priori. Las probabilidades a priori, o ponderaciones de clase, especifican la probabilidad de que un caso caiga en cada una de las clases de la variable dependiente, sin tener ningún conocimiento previo de los valores de los predictores. Las probabilidades a priori son parte fundamental de cualquier árbol de decisión y la mayoría del software actual permite utilizar ponderaciones estimadas según las proporciones de cada clase, aunque no siempre sea el camino óptimo. Así mismo, el software actual ofrece adicionalmente la posibilidad de tratar las clases como si fueran del mismo tamaño, especificando probabilidades iguales para cada clase. También algunos programas permiten utilizar matrices activas de costes, que permiten que el árbol de decisión se vaya adaptando en cada uno de los nodos para evitar los mayores costes.

©Alfaomega-RC Libros

Selección del método de división. Se trata de elegir el método con el que seleccionar, en cada uno de los niveles del proceso de división, la mejor división posible del mejor predictor. En la actualidad predominan fundamentalmente los enfoques mediante métodos exhaustivos y métodos de tipo discriminante. En cuanto a los *métodos exhaustivos*, el más conocido y simple consiste en examinar todas las posibles divisiones de los datos según cada predictor y seleccionar la división que produce clasificaciones más puras (observando la mejoría en la bondad de ajuste a través de una serie de medidas como Gini, entropía, $\chi 2$, twoing, symgini, twoing ordenado, desviación de mínimos cuadrados y combinaciones lineales). Este método lo utilizan CART y CHAID exhaustivo. En cuanto a los *métodos de tipo discriminante*, se sigue un proceso distinto y computacionalmente más sencillo. En vez de buscar a la vez la mejor variable y su mejor punto de división, se abordan estos dos problemas por separado. En cada nodo, calculan primero un test $\chi 2$ (para cada predictor categórico) o un ANOVA (para cada predictor métrico), seleccionándose de entre todas las variables significativas, la que proporciona probabilidades asociadas menores. En una segunda fase, se aplica un análisis discriminante sobre el predictor con el fin de encontrar la mejor división posible de la variable. Estos procedimientos son utilizados en los árboles QUEST.

Elección del tamaño adecuado o *problema del sobreajuste*. Si no se establece ningún límite en el número de divisiones a ejecutar en un árbol, se corre el riesgo de encontrarnos con muy pocos elementos en cada clase y un número ingente de divisiones que provocan sobreajuste. Para tratar este problema se utilizan las reglas de parada y la poda. Las *reglas de parada* detienen la generación de nuevas divisiones cuando estas supongan una mejora muy pequeña de la predicción. Entre las reglas de parada directa para detener automáticamente el proceso de construcción del árbol, tenemos la extensión máxima del árbol o número de niveles máximos permitidos por debajo del nodo raíz, el mínimo número de casos en un nodo que acota el número de nodos, impidiendo que no sobrepasen un número determinado de casos y la mínima fracción de objetos, mediante la cual los nodos no contendrán más casos que una fracción determinada del tamaño de una o más clases. La regla de parada la establece a priori el propio investigador, en función de investigaciones pasadas, análisis previos, o incluso en función de su propia experiencia e intuición. En cuanto a la *poda*, existe siempre el riesgo de no descubrir estructuras relevantes en los datos debido a una finalización prematura del análisis. Por ello, se sugiere un enfoque alternativo en dos fases. En una primera fase se desarrolla un enorme árbol que contenga cientos o incluso miles de nodos. En una segunda fase, el árbol es podado, eliminándose las ramas innecesarias hasta dar con el tamaño adecuado del árbol. Este proceso automático y retrospectivo, que compara simultáneamente todos los posibles subárboles resultado de podar en

diferente grado el árbol original, no debe confundirse con la opción que ofrecen algunos programas (particularmente los de tipo CHAID) de podar manualmente el árbol una vez que se ha llegado a la solución final, opción esta que no elimina los problemas de utilizar reglas de parada.

TIPOS DE ÁRBOLES DE DECISIÓN

Los tres tipos de árboles más utilizados hoy en día son: los árboles CHAID, los árboles CART y los árboles QUEST.

Árboles CHAID

El método CHAID (*Chi-square Automatic Interaction Detector*) es la conclusión de una serie de métodos basados en el detector Automático de Interacciones (AID) de Morgan y Sonquist. Se trata de un método exploratorio de análisis de datos, útil para identificar variables importantes, y sus interacciones enfocadas a la segmentación y a los análisis descriptivos, que suelen ser pasos previos a otros análisis posteriores. La variable dependiente puede ser cualitativa (nominal u ordinal) o cuantitativa. Para variables cualitativas, el análisis lleva a cabo una serie de análisis $\chi 2$ entre las variables dependiente y predictora. En el caso de variables dependientes cuantitativas, se recurre a métodos de análisis de varianza, en los que los intervalos (divisiones) se determinan óptimamente para las variables independientes, de forma que maximicen la capacidad para explicar la varianza de la medida dependiente. Se divide cada nodo localizando el par de categorías permisible del predictor con el menor valor de $\chi 2$. Si el nivel de significación es menor que un cierto nivel crítico, se unen ambas categorías y se repite el proceso. Si es mayor, se convierten en dos candidatas a la división de la variable. Este proceso continúa con cada par de categorías, hasta que dejan de producirse uniones y posibles divisiones. La última candidata a la división (que generalmente no suele coincidir con la división más significativa) es la que se elige para dividir al predictor. El proceso se repite de forma recursiva en cada uno de los nodos, hasta que se activa cualquiera de las reglas de parada del proceso.

Este método ahorra bastante tiempo de computación, pero no garantiza que sea capaz de encontrar realmente la mejor división posible en cada modo.

Para garantizar el hallazgo de la división más significativa se utiliza el *método CHAID exhaustivo*, que trata a todas las variables por igual, independientemente del tipo de variable y del número de categorías. Por otro lado, este método permite trabajar con variables dependientes categóricas y métricas. Las variables categóricas

 ©Alfaomega-RC Libros

utilizan el estadístico $\chi 2$ y dan lugar a un *árbol de clasificación*. Las variables métricas utilizan el estadístico F y dan lugar a lo que se conoce como *árboles de regresión*. También permite utilizar predictores de tipo métrico, mediante su conversión previa en variables categóricas. Los métodos CHAID producen divisiones de la validación cruzada en más de dos grupos, lo cual siempre es un valor añadido.

Árboles CART

El método CART (*Classification And Regression Trees*) o C&RT es una alternativa al CHAID exhaustivo para *árboles de clasificación* (variables dependientes categóricas). Este método nació para intentar superar algunas de las deficiencias y debilidades que por entonces mostraba la formulación original del CHAID, que estaba limitado inicialmente a variables dependientes nominales y variables independientes categóricas hasta la aparición de su versión exhaustiva. Estaba claro que se necesitaba utilizar predictores de cualquier nivel de medida. Además, CART tiene una estructura estadística más fuerte que CHAID, lo que le llevó a ser utilizado en campos de la investigación como la medicina, además de en el marketing. CART se utiliza para árboles de clasificación con variable dependiente cualitativa y para árboles de regresión con variable dependiente cuantitativa, y genera árboles binarios.

El método comienza dividiendo la muestra en subconjuntos y evaluando cada predictor cuantitativo para encontrar el mejor punto de corte o cada predictor categórico y para encontrar las mejores agrupaciones de categorías. A continuación se comparan también los predictores, seleccionándose el predictor y la división que produce la mayor bondad de ajuste. Para predictores cuantitativos suele utilizarse la minimización del error cuadrático o de la desviación media absoluta respecto de la mediana. Para predictores cualitativos suele utilizarse el coeficiente Gini para evaluar la probabilidad de una mala clasificación (valor cero para clasificación perfecta y valor uno para una mala clasificación).

No debemos de olvidar que los métodos CHAID producen divisiones de la validación cruzada en más de dos grupos, mientras que el método CART solo produce divisiones binarias.

Árboles QUEST

Los árboles QUEST (*Quick, Unbiased, Efficient, Statistical Tree*) consisten en un algoritmo de clasificación arborescente creado específicamente para solventar dos de los principales problemas que presentan métodos como CART y CHAID

exhaustivo, a la hora de dividir un grupo de sujetos en función de una variable independiente. Este tipo de árboles mitigan la complejidad computacional (enfoque de cálculo más sencillo) y los sesgos en la selección de variables. Se trata de evitar que se seleccionen aquellas variables que cuentan con un mayor número de categorías. QUEST intenta seleccionar el mejor predictor y su mejor punto de corte como tareas separadas, calculando en cada nodo la asociación entre cada predictor y la variable dependiente mediante el estadístico F del ANOVA o la F de Levene para predictores continuos y ordinales o mediante una $\chi 2$ de Pearson para predictores nominales. Se consiguen divisiones binarias de la variable dependiente mediante la creación de dos superclases en el predictor, aplicando un algoritmo conglomerativo.

Por último, para eliminar el sesgo en la selección de variables, se elige el predictor que tiene la mayor asociación con la variable dependiente. Posteriormente, para hallar el mejor punto de corte se recurre a un análisis discriminante cuadrático, repitiéndose el proceso recursivamente hasta que lo permitan las reglas de parada establecidas en el algoritmo. De esta forma, se eliminan sesgos de respuesta y se simplifica el cálculo.

En cuanto a la valoración de los métodos de construcción de árboles, podría establecerse un orden de jerarquía (nunca absoluto) que sitúe el método QUEST como superior a CART y este último método superior a CHAID. No olvidemos que QUEST admite métodos de validación mediante poda y permite utilizar combinaciones lineales de variables. Pero debe quedar claro que esta evaluación solo es válida en líneas generales.

ÁRBOLES DE DECISIÓN CON IBM SPSS MODELER

Dentro de la categoría *Clasificación* de la ficha *Modelado*, IBM SPSS Modeler dispone de los nodos *C5.0, Árbol C&R, Quest y CHAID* cuya finalidad es la realización de árboles de decisión y que se muestran en la Figura 7-25.

Figura 7-25

- *C5.0*: permite construir árboles de decisión y conjunto de reglas utilizando el algoritmo C5.0.

- *Árboles de clasificación y regresión (Árbol C&R):* permite construir árboles de decisión para clasificación y regresión con la finalidad de predecir.

©Alfaomega-RC Libros

- *Quest*: permite construir árboles utilizando el algoritmo Quest.

- CHAID: permite construir árboles utilizando el algoritmo CHAID.

El nodo C5.0

El nodo *Crear C5.0* de la categoría *Clasificación* de la ficha *Modelado* se utiliza para construir árboles de decisión y conjunto de reglas utilizando el algoritmo C5.0. Como ejemplo, vamos a utilizar los datos del archivo *creditos1.txt* para establecer las reglas lógicas que van a permitir ayudar a tomar la decisión sobre cuándo decidir asignar o no un crédito. Una vez situado el fichero ASCII rellenando la pantalla de *Archivo variable*, a través del nodo *Tipo* asignaremos a la variable *CREDIT_V* el papel de variable objetivo (SALIDA) y el resto de variables serán *inputs* (ENTRADA) excepto cliente que no se utilizará (Figura 7-26). Añadimos al diagrama el nodo *C5.0* y hacemos clic sobre él con el botón derecho del ratón para elegir la opción *Edición*. Obtenemos la pantalla del árbol de decisión en cuya solapa *Campo* elegimos la opción *Utilizar la configuración del nodo Tipo* (Figura 7-27). En la solapa *Modelo* elegimos *Automático, Árbol de decisión* y *Experto* utilizando una poda del 75% (Figura 7-28). En la solapa *Analizar* se elige *Calcular la importancia del predictor* y *Calcular puntuaciones brutas de propensión* (Figura 7-38). Al hacer clic en *Aplicar* y *Ejecutar* se obtiene el modelo de árbol C5.0 en la ruta (Figura 7-39).

Si hacemos clic con el botón derecho del ratón sobre el icono del modelo en la ruta y elegimos *Edición*, se obtiene la pantalla de resultados del árbol cuya paleta *Modelo* presenta la importancia de los predictores (Figura 7-29). PAGO_MES es el predictor más importante en la decisión de conceder créditos. En la paleta *Visor* se observa la estructura del árbol (Figura 7-30) y en la paleta *Resumen* se observa un resumen del modelo (Figura 7-31).

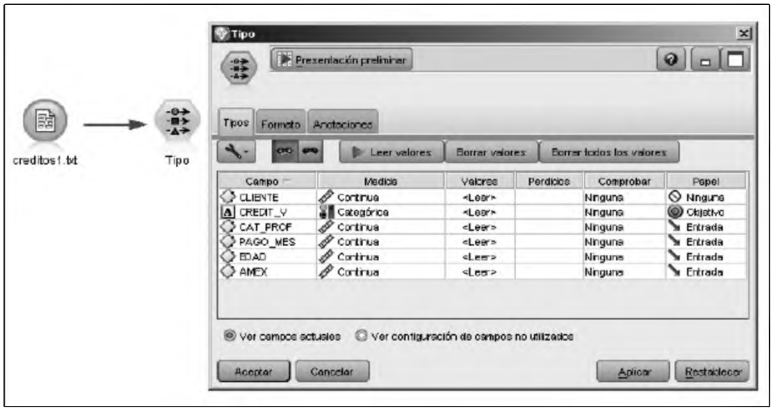


Figura 7-26

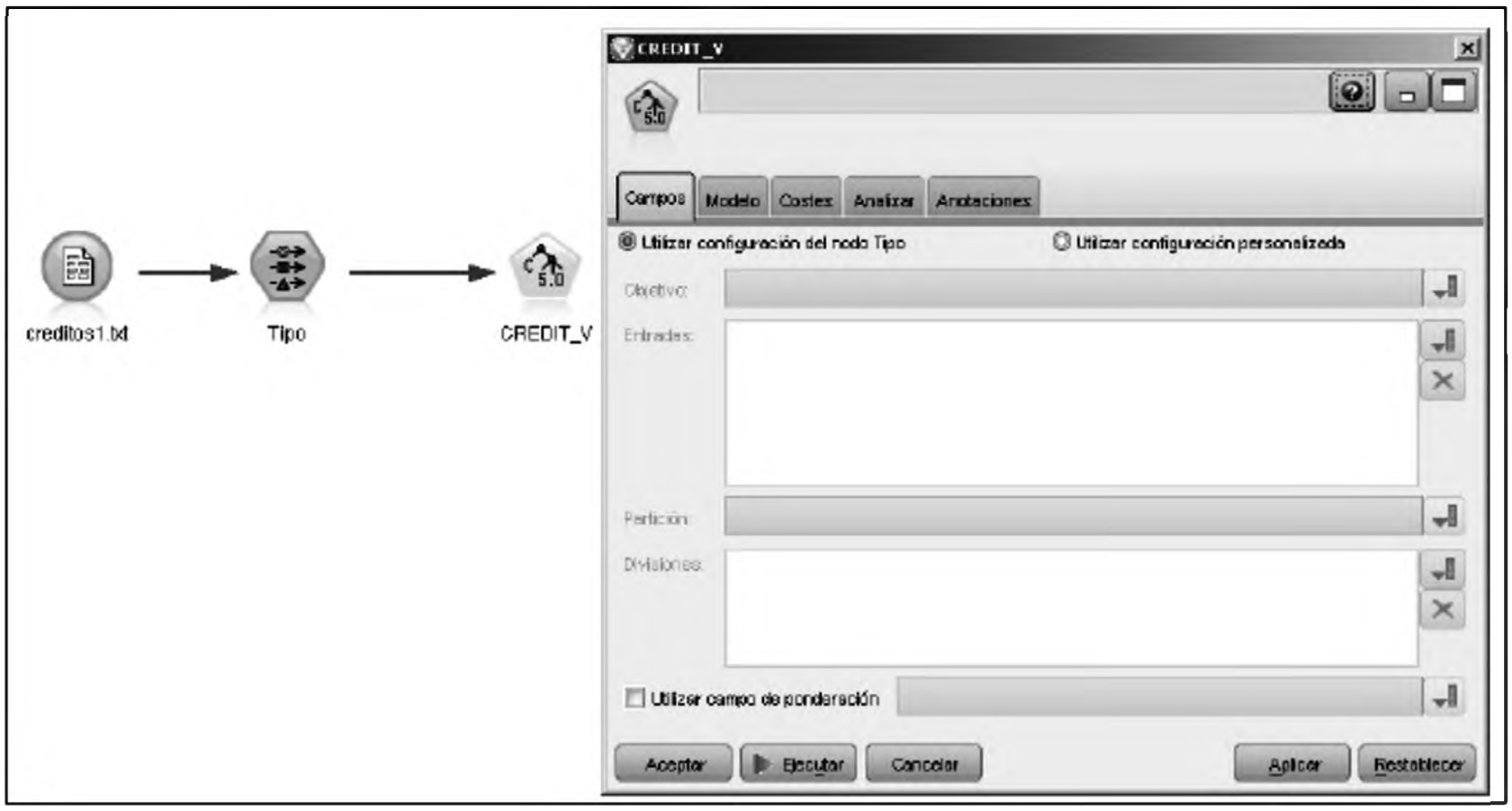


Figura 7-27

Figura 7-28

Figura 7-29

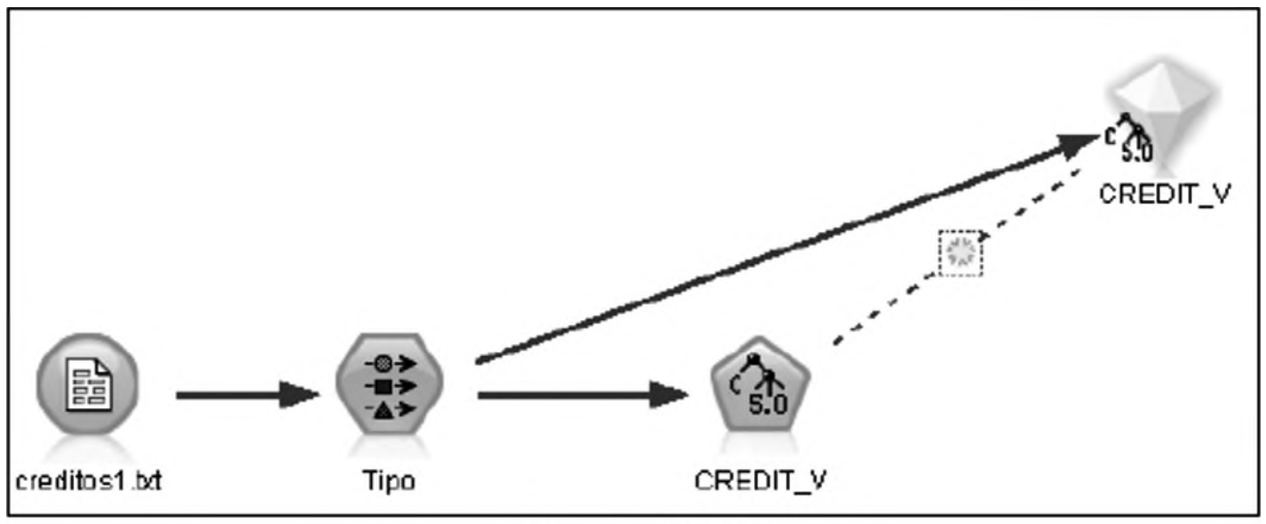


Figura 7-30

©Alfaomega-RC Libros

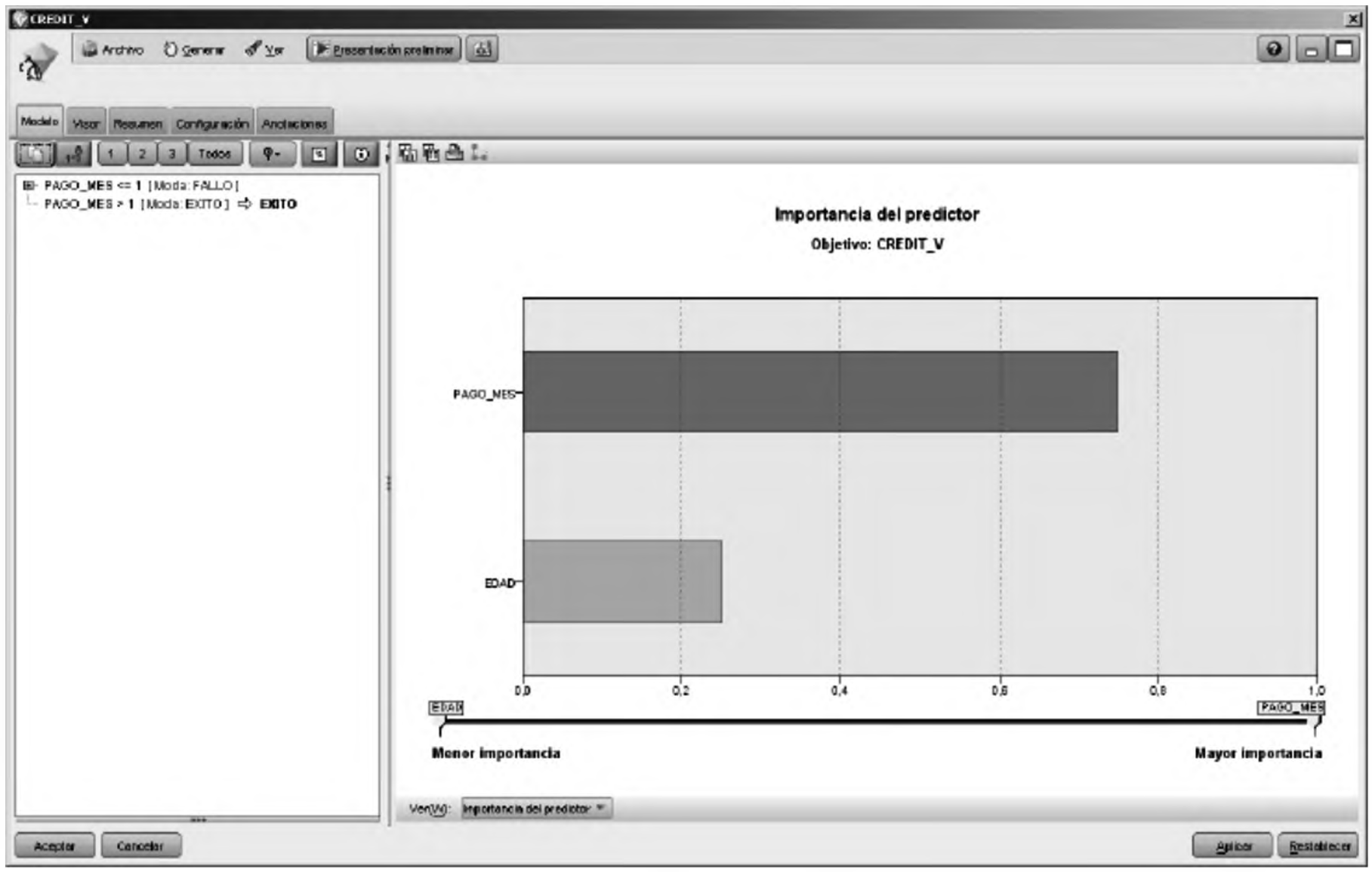


Figura 7-31

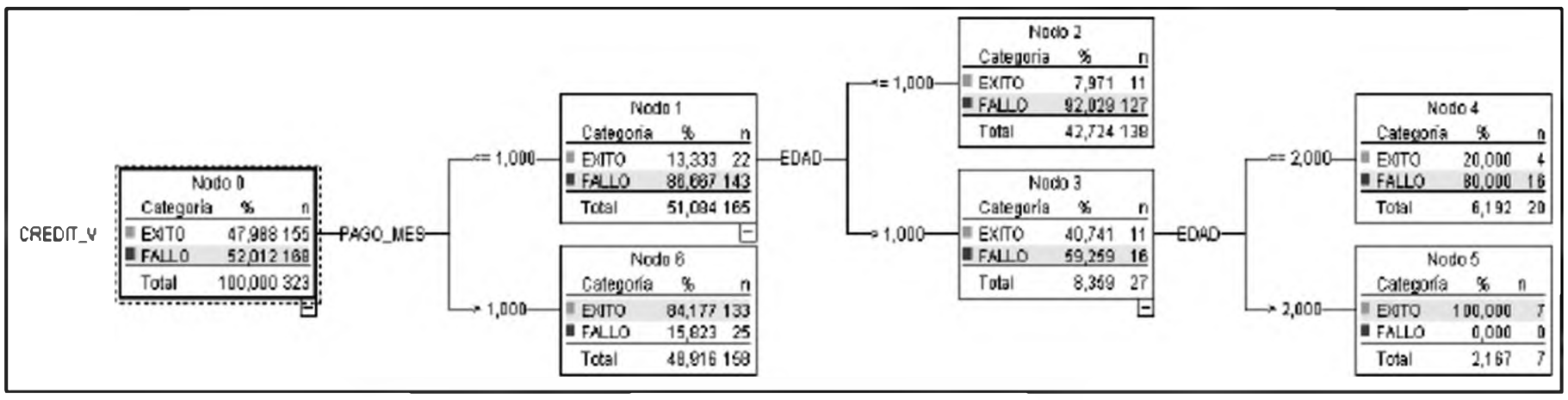


Figura 7-32

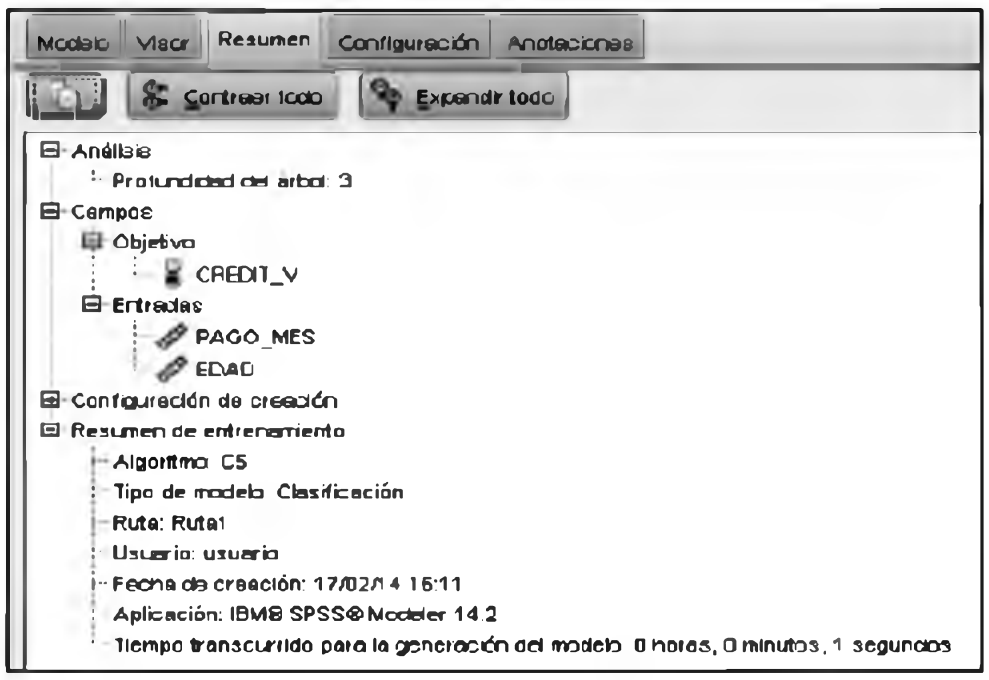


Figura 7-33

El nodo Árbol C&R

El nodo *Árbol C&R* de la paleta *Modelado* se utiliza para construir árboles de decisión para clasificación y regresión con la finalidad de predecir. Como ejemplo, vamos a utilizar los datos del archivo *creditos1.txt* para establecer las reglas lógicas que van a permitir ayudar a tomar la decisión sobre cuándo decidir asignar o no un crédito. Una vez situado el fichero ASCII rellenando la pantalla de *Archivo variable*, a través del nodo *Tipo* asignaremos a la variable *CREDIT_V* el papel de variable objetivo (SALIDA) y el resto de variables serán *inputs* (ENTRADA) excepto cliente que no se utilizará (Figura 7-34). Añadimos al diagrama el nodo *Árbol C&R* y hacemos clic sobre él con el botón derecho del ratón para elegir la opción *Edición*. Obtenemos la pantalla del árbol de decisión en cuya solapa *Campo* elegimos las opciones de la Figura 7-35. En la solapa *Opciones de generación* en su categoría *Objetivo* elegimos las opciones de la Figura 7-36. En las categorías *Procedimientos básicos, Reglas de parada, Costes y previas, Conjuntos* y *Avanzado* elegimos las opciones por defecto. En la solapa *Opciones del modelo* elegimos *Automático* como se indica en la Figura 7-37. Al hacer clic en *Aplicar* y *Ejecutar* se obtiene el modelo de árbol *C$R* en la ruta (Figura 7-39).

Si hacemos clic con el botón derecho del ratón sobre el icono del modelo en la ruta y elegimos *Edición*, se obtiene la pantalla de resultados del árbol cuya paleta *Modelo* presenta la importancia de los predictores (Figura 7-39). PAGO_MES es el predictor más importante en la decisión de conceder créditos. Se observa que ahora se tienen en cuenta más predictores significativos que en el árbol C 5.0. En la paleta *Visor* se observa la estructura del árbol (Figura 7-40).

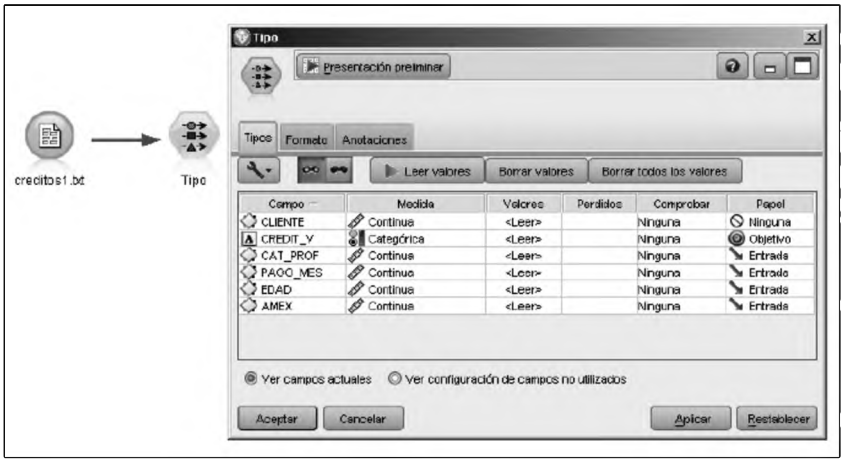


Figura 7-34

©Alfaomega-RC Libros

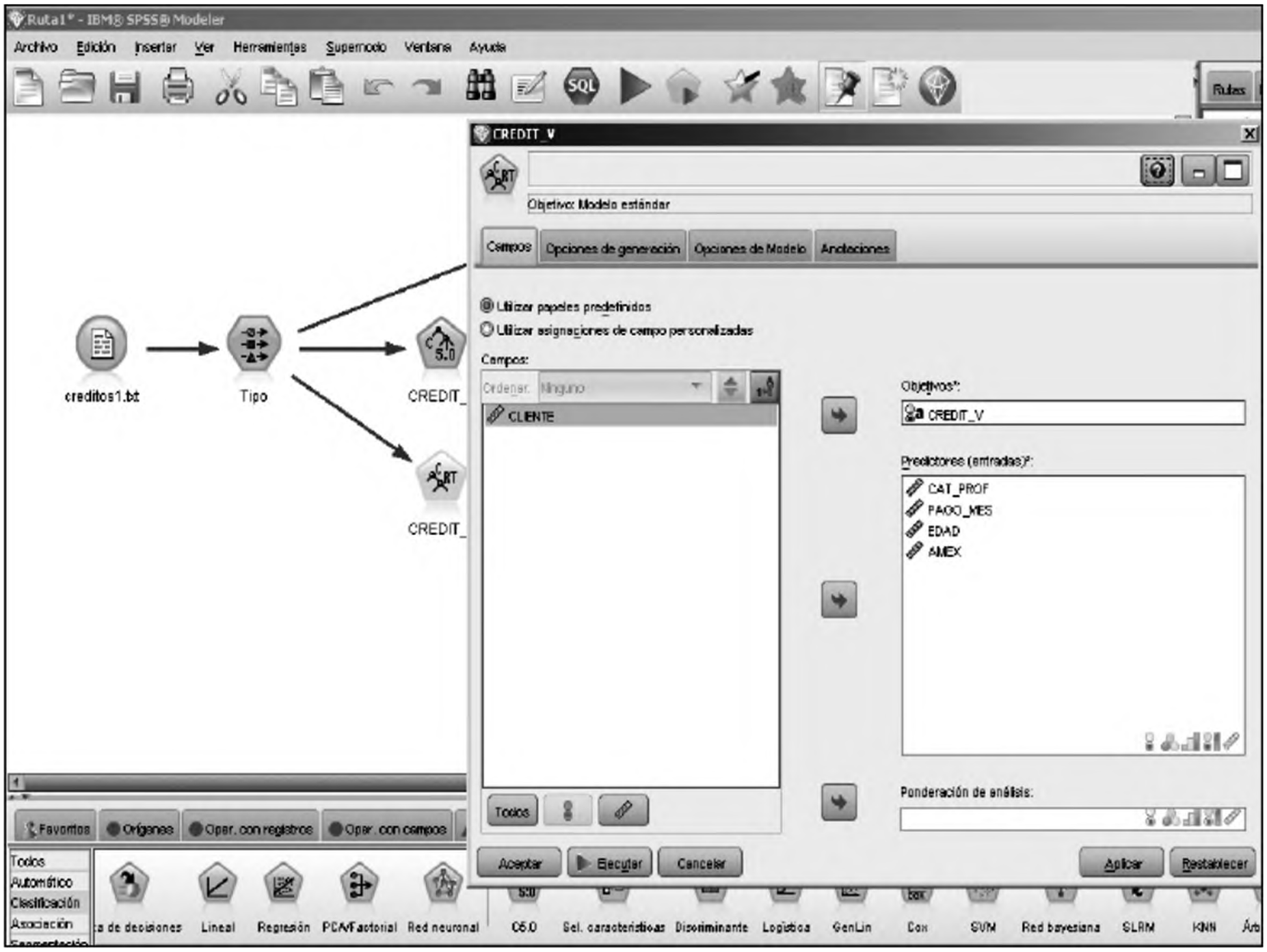


Figura 7-35

Figura 7-36

Figura 7-37

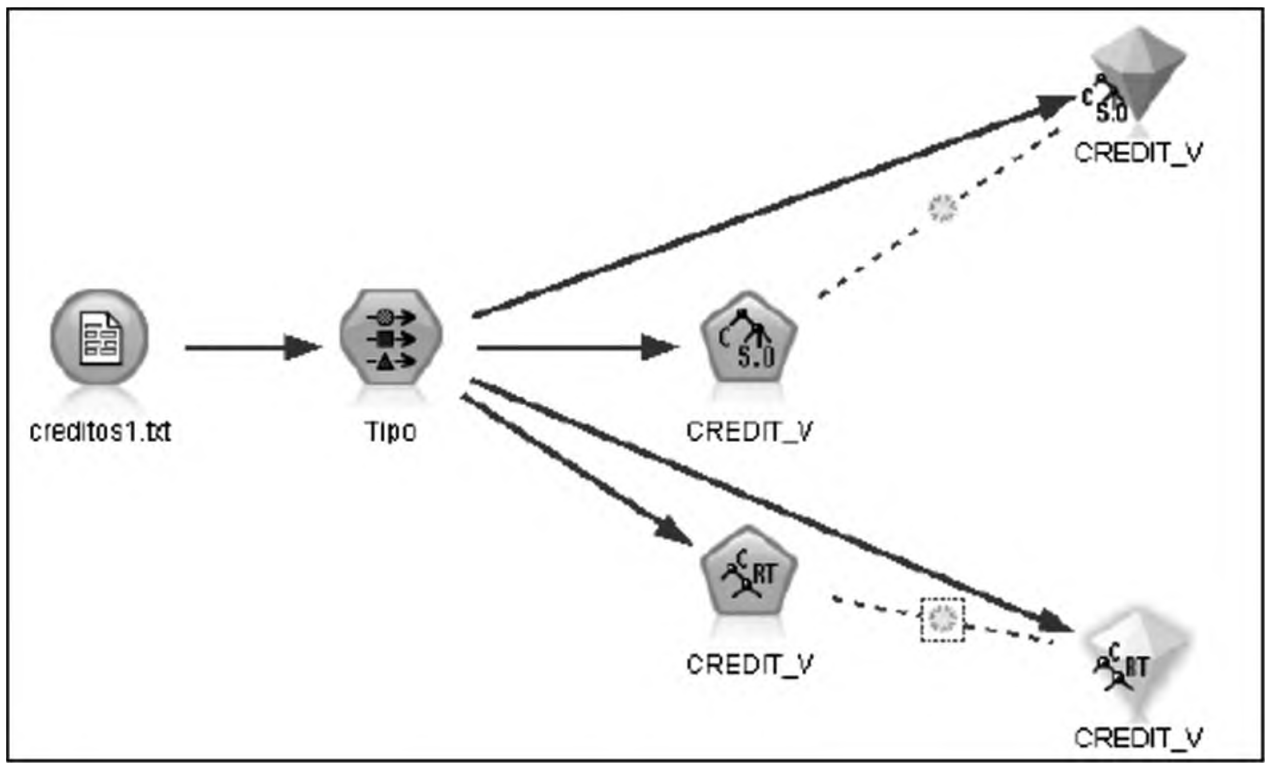


Figura 7-38

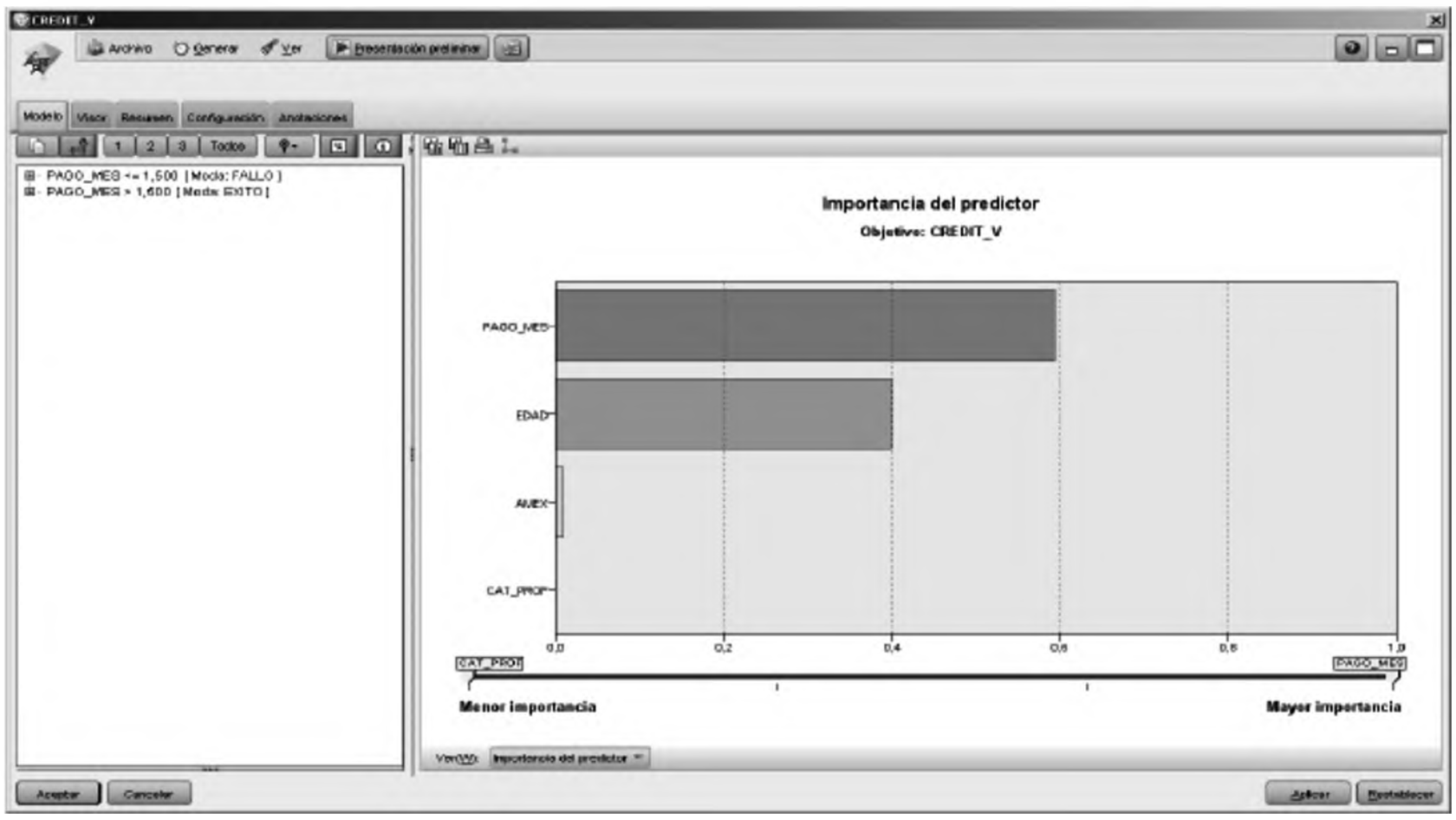


Figura 7-39

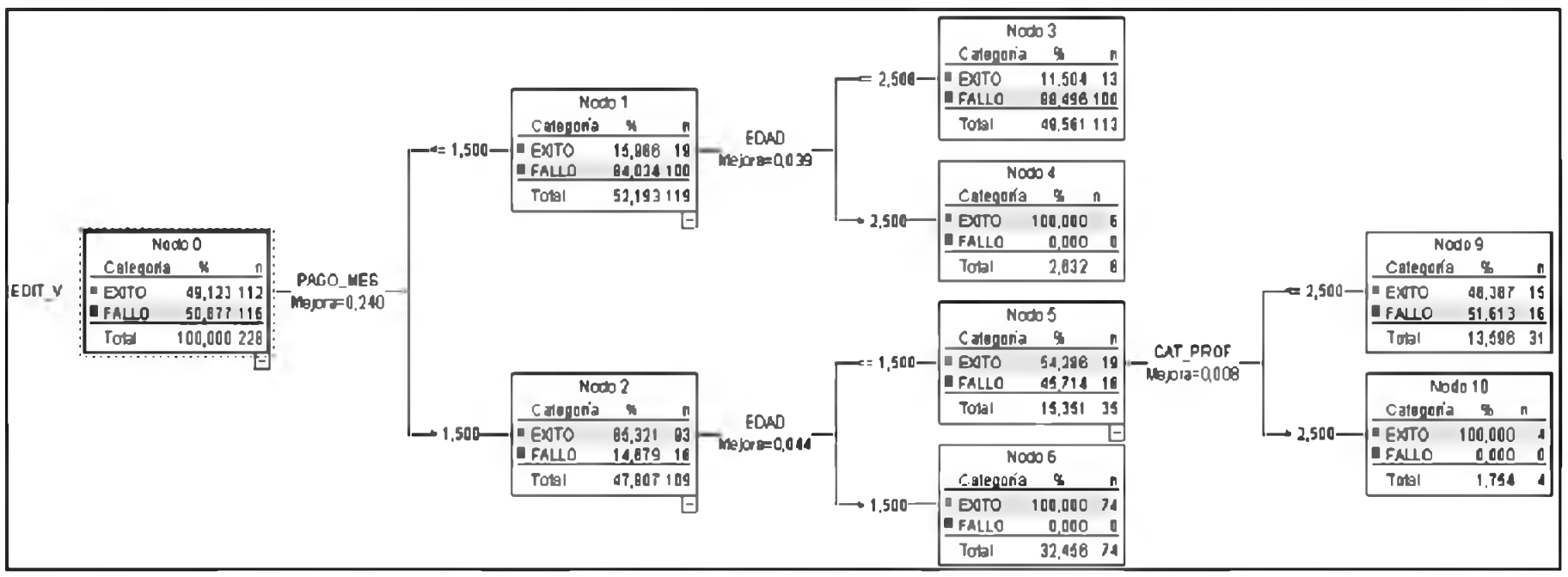


Figura 7-40

©Alfaomega-RC Libros

El nodo Quest

El nodo *Quest* de la paleta *Modelado* se utiliza para construir árboles de decisión mediante el algoritmo QUEST. Como ejemplo, vamos a utilizar los datos del archivo *creditos1.txt* para establecer las reglas lógicas que van a permitir ayudar a tomar la decisión sobre cuándo decidir asignar o no un crédito. Una vez situado el fichero ASCII rellenando la pantalla de *Archivo variable*, a través del nodo *Tipo* asignaremos a la variable *CREDIT_V* el papel de variable objetivo (SALIDA) y el resto de variables serán *inputs* (ENTRADA) excepto cliente que no se utilizará (Figura 7-41). Añadimos al diagrama el nodo *Quest* y hacemos clic sobre él con el botón derecho del ratón para elegir la opción *Edición*. Obtenemos la pantalla del árbol de decisión en cuya solapa *Campo* elegimos las opciones de la Figura 7-42. En la solapa *Opciones de generación* en su categoría *Objetivo* elegimos las opciones de la Figura 7-43. En las categorías *Procedimientos básicos, Reglas de parada, Costes y previas, Conjuntos* y *Avanzado* elegimos las opciones por defecto. En la solapa *Opciones del modelo* elegimos *Automático* como se indica en la Figura 7-44. Al hacer clic en *Aplicar* y *Ejecutar* se obtiene el modelo de árbol *Quest* en la ruta (Figura 7-45).

Si hacemos clic con el botón derecho del ratón sobre el icono del modelo en la ruta y elegimos *Edición*, se obtiene la pantalla de resultados del árbol cuya paleta *Modelo* presenta la importancia de los predictores (Figura 7-46). PAGO_MES es el predictor más importante en la decisión de conceder créditos. Se observa que ahora se tienen en cuenta otros predictores significativos que no coinciden con los de los otros árboles ya vistos. En la paleta *Visor* se observa la estructura del árbol (Figura 7-47).

Figura 7-41

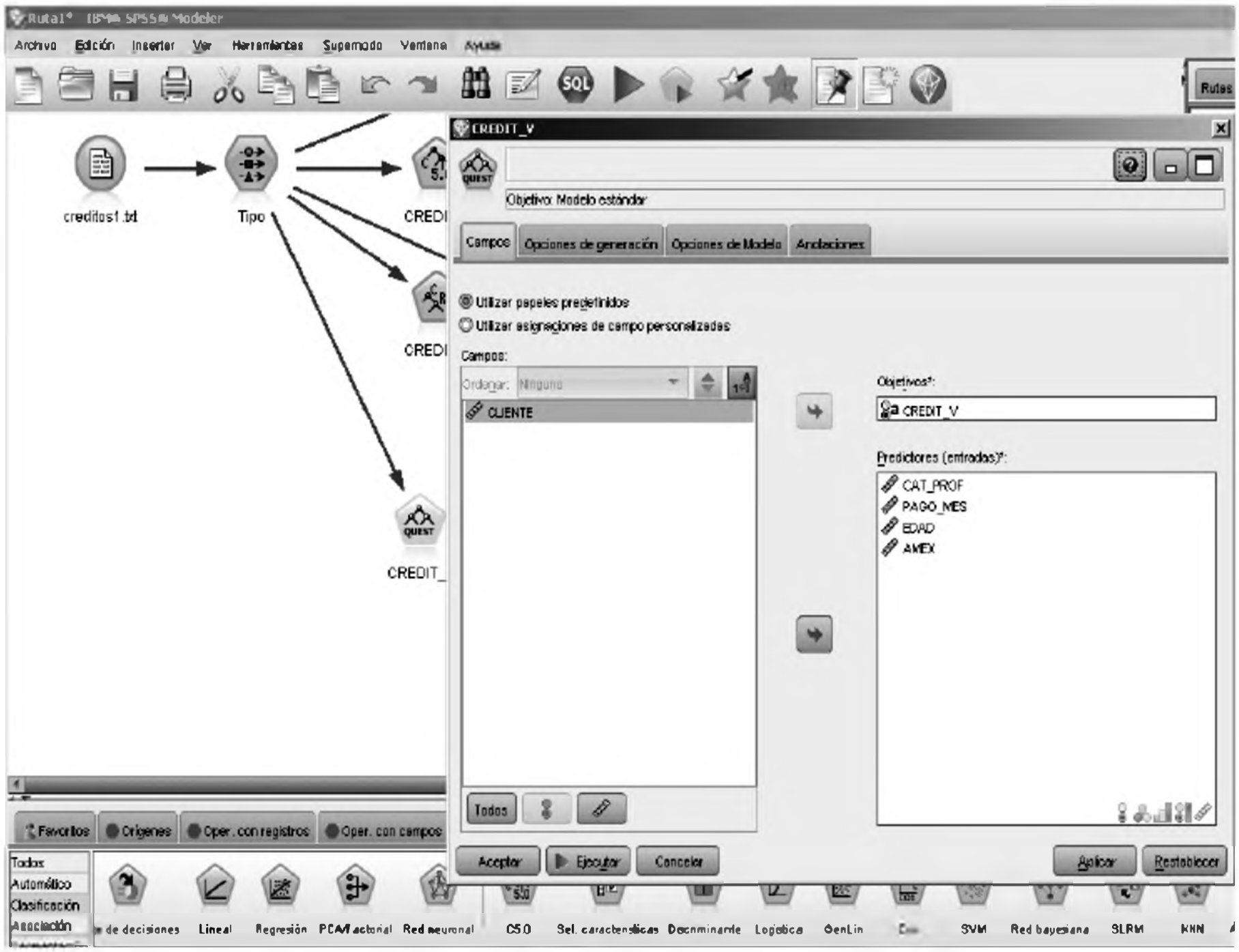


Figura 7-42

Figura 7-43

Figura 7-44

©Alfaomega-RC Libros

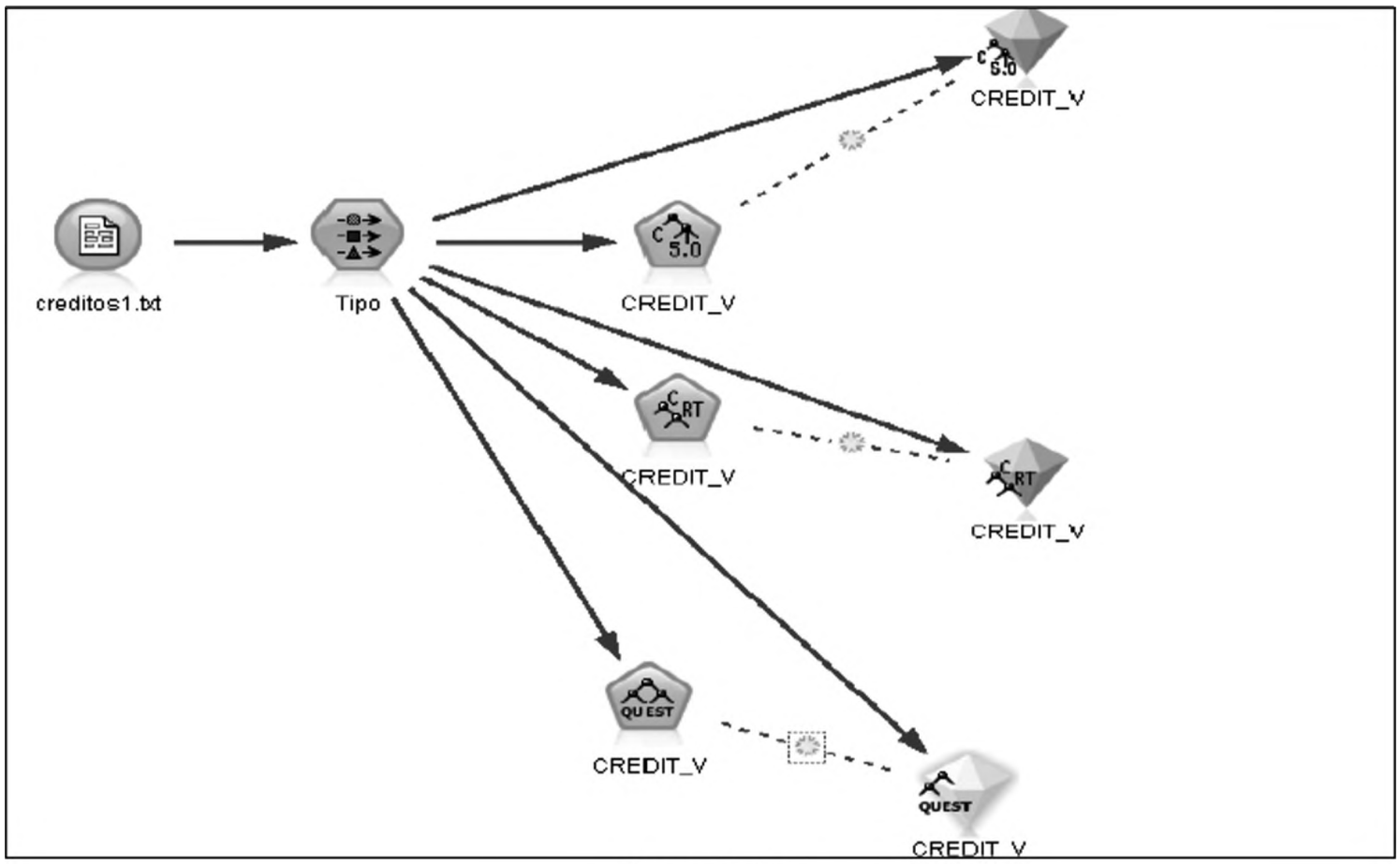


Figura 7-45

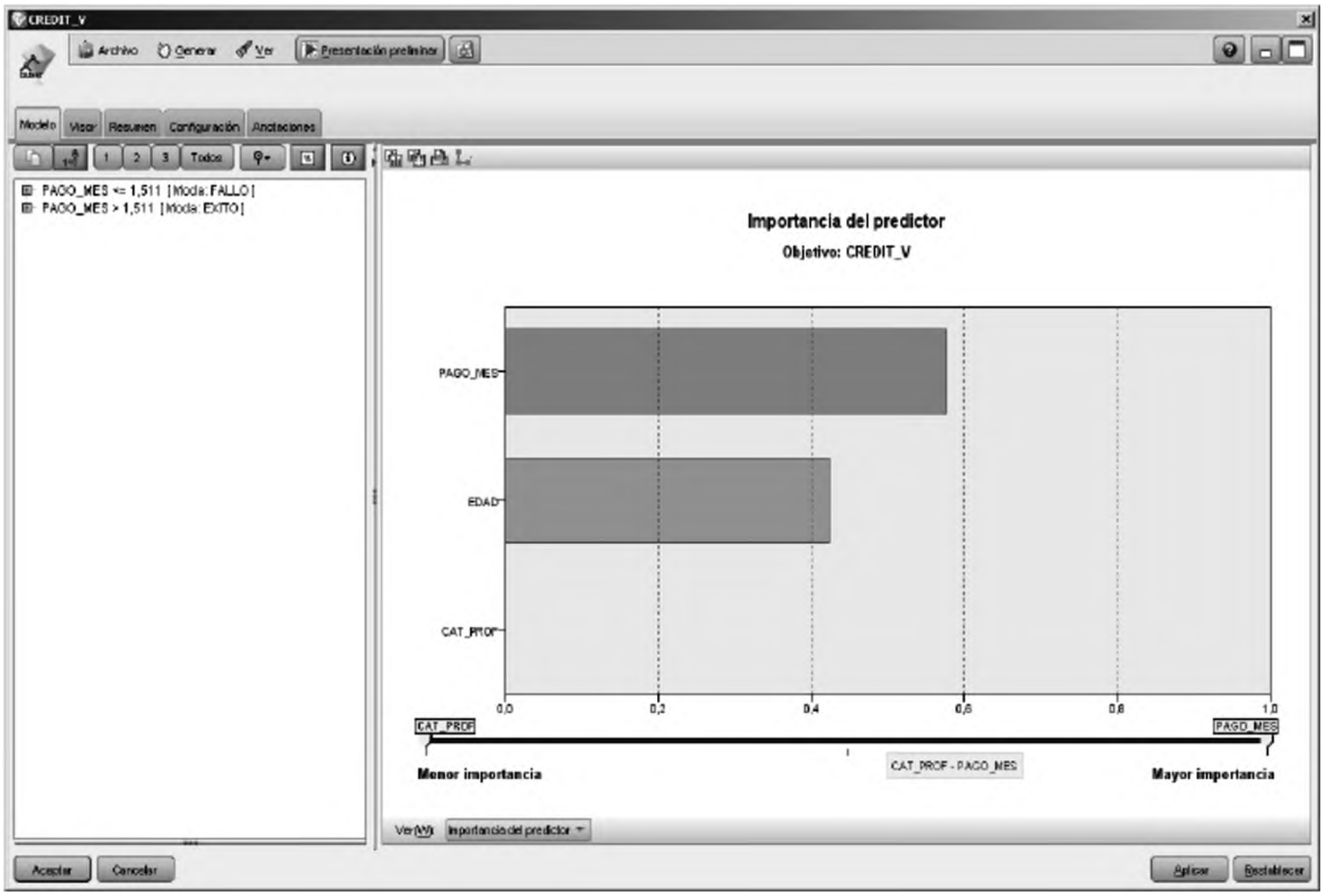


Figura 7-46

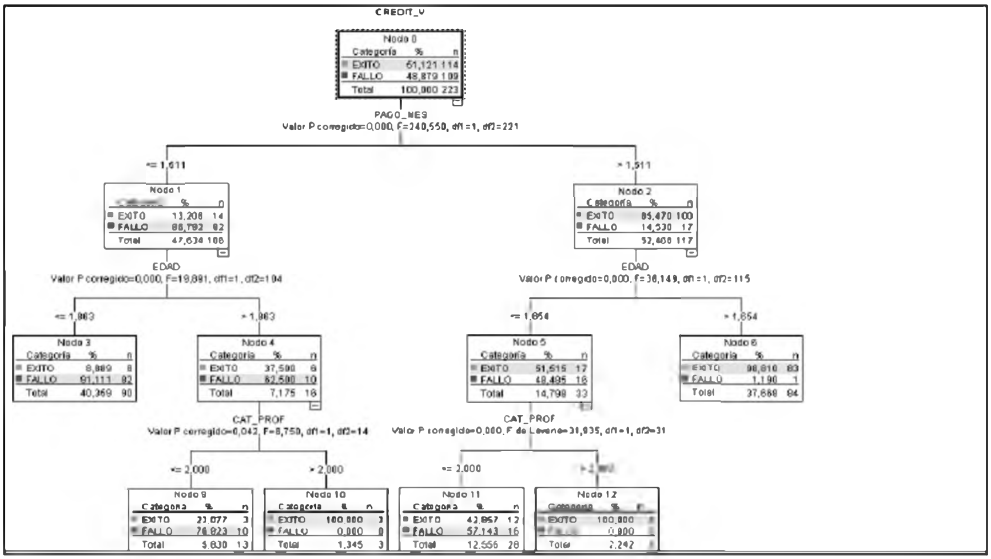


Figura 7-47

El nodo CHAID

El nodo *CHAID* de la paleta *Modelado* se utiliza para construir árboles de decisión mediante el algoritmo CHAID. Como ejemplo, vamos a utilizar los datos del archivo *creditos1.txt* para establecer las reglas lógicas que van a permitir ayudar a tomar la decisión sobre cuándo decidir asignar o no un crédito. Una vez situado el fichero ASCII rellenando la pantalla de *Archivo variable*, a través del nodo *Tipo* asignaremos a la variable *CREDIT_V* el papel de variable objetivo (SALIDA) y el resto de variables serán *inputs* (ENTRADA) excepto cliente que no se utilizará (Figura 7-48). Añadimos al diagrama el nodo *CHAID* y hacemos clic sobre él con el botón derecho del ratón para elegir la opción *Edición*. Obtenemos la pantalla del árbol de decisión en cuya solapa *Campo* elegimos las opciones de la Figura 7-49. En la solapa *Opciones de generación* en su categoría *Objetivo* elegimos las opciones de la Figura 7-50. En las categorías *Procedimientos básicos, Reglas de parada, Costes y previas, Conjuntos* y *Avanzado* elegimos las opciones por defecto. En la solapa *Opciones del modelo* elegimos *Automático* como se indica en la Figura 7-51. Al hacer clic en *Aplicar* y *Ejecutar* se obtiene el modelo de árbol *CHAID* en la ruta (Figura 7-52). Si hacemos clic con el botón derecho del ratón sobre el icono del modelo en la ruta y elegimos *Edición*, se obtiene la pantalla de resultados del árbol cuya paleta *Modelo* presenta la importancia de los predictores (Figura 7-53). PAGO_MES es el predictor más importante en la decisión de conceder créditos. Se observa que ahora se tienen en cuenta otros predictores significativos que no coinciden con los de los otros árboles ya vistos. En la paleta *Visor* se observa la estructura del árbol (Figura 7-54).

 ©Alfaomega-RC Libros

Figura 7-48

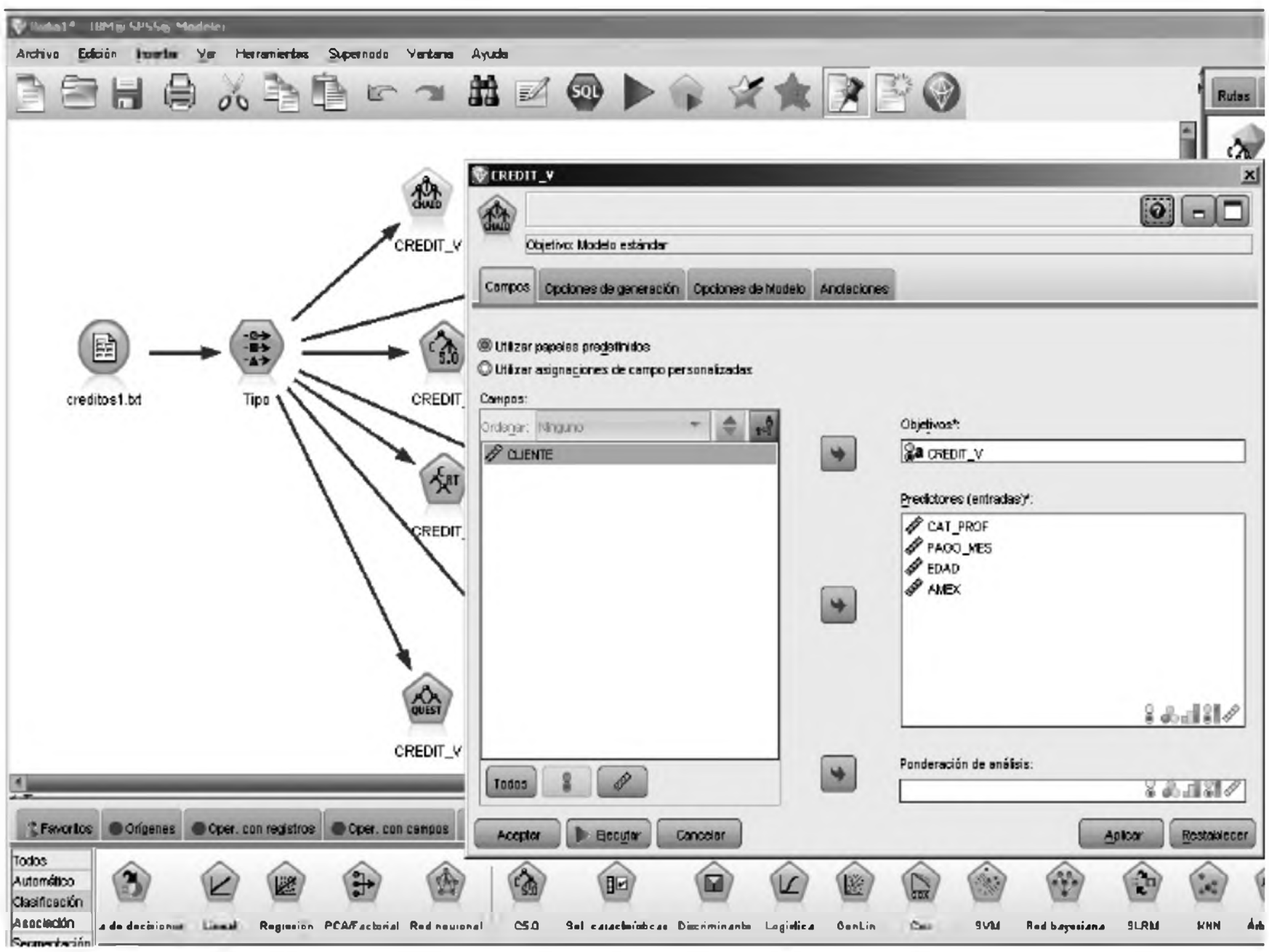


Figura 7-49

Figura 7-50

Figura 7-51

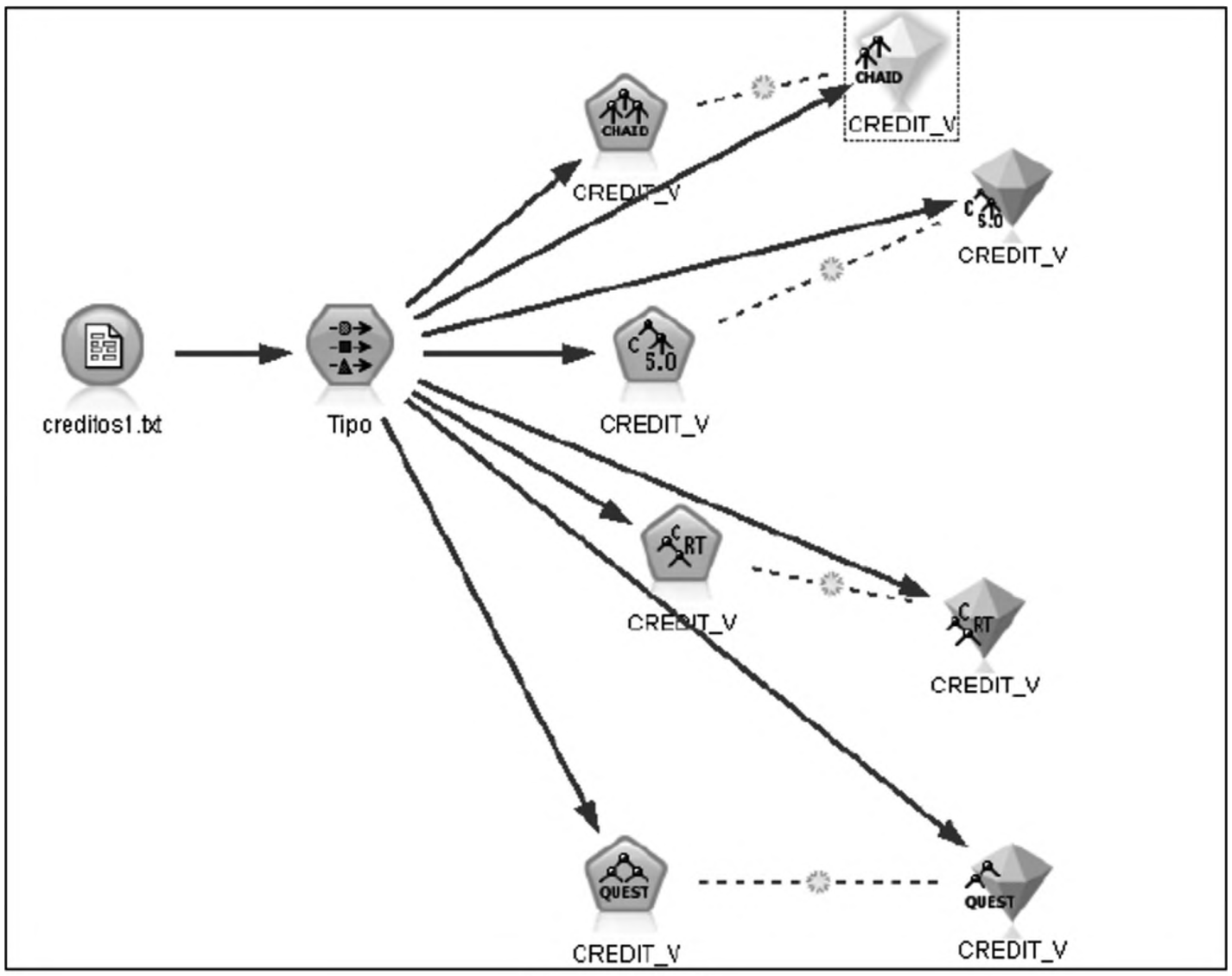


Figura 7-52

©Alfaomega-RC Libros

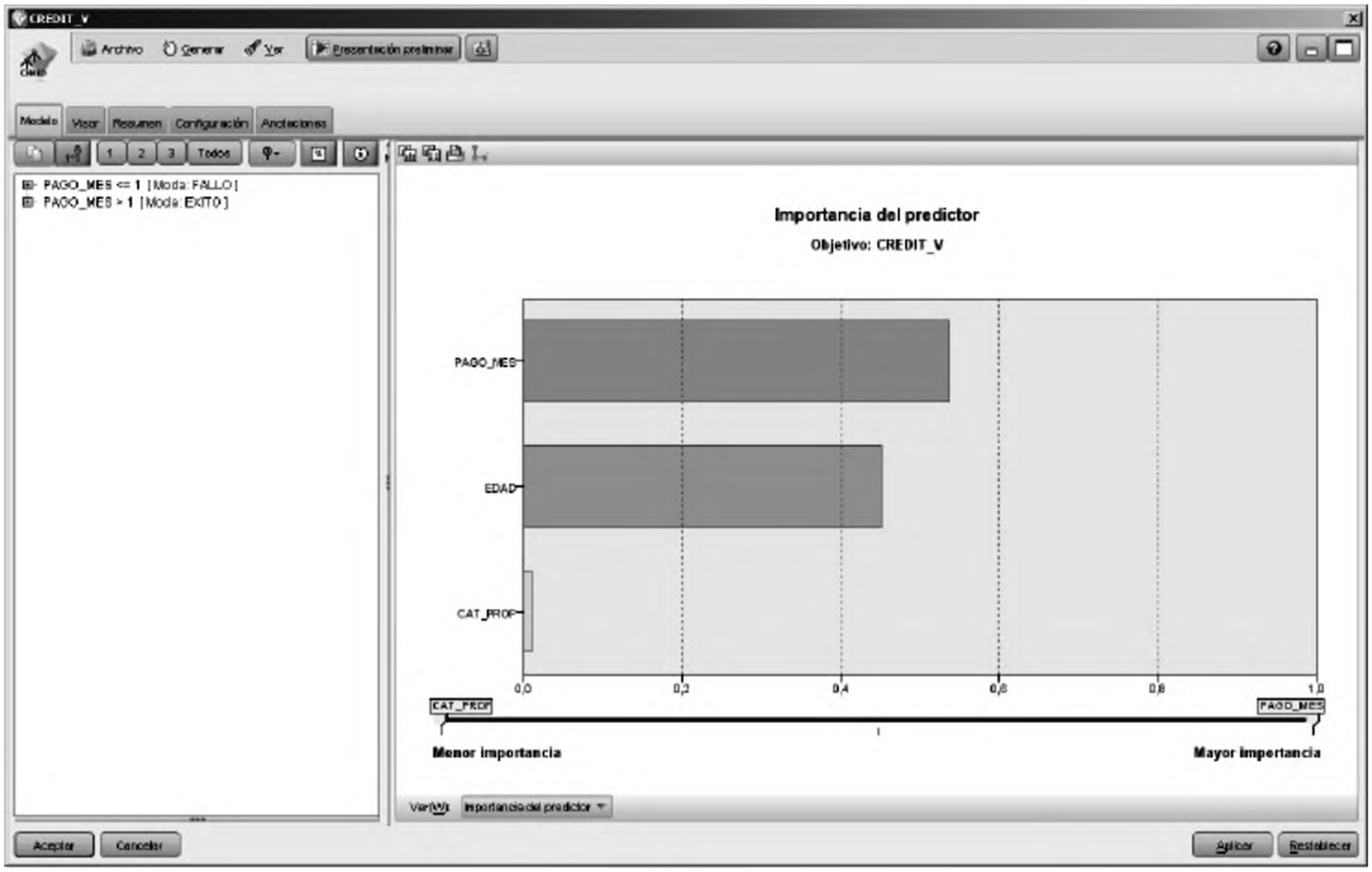


Figura 7-53

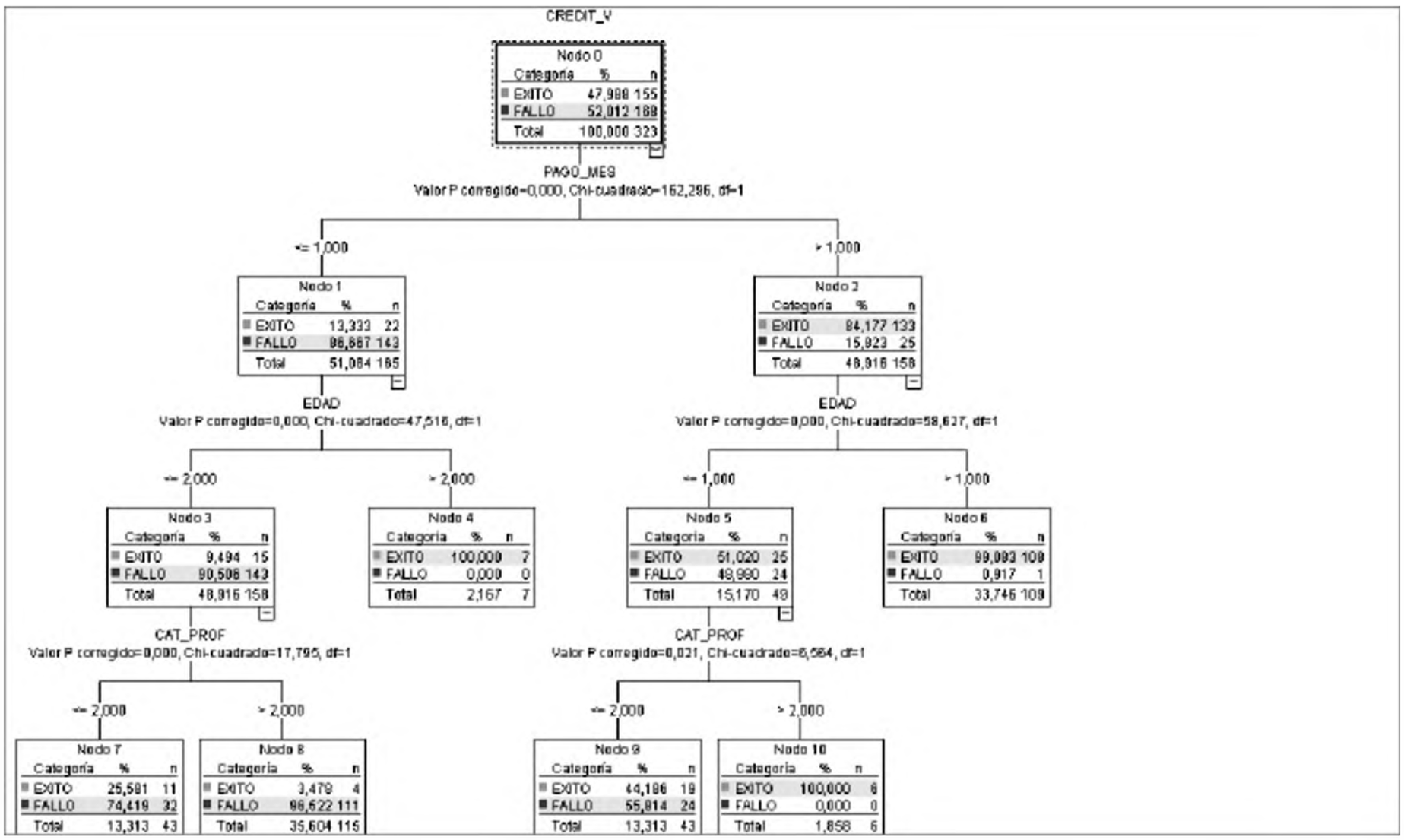


Figura 7-54

CAPÍTULO 8

SEGMENTACIÓN Y PERFILADO CON HERRAMIENTAS DE SAS

ANÁLISIS CLUSTER CON SAS ENTERPRISE MINER. EL NODO CLUSTERING

El nodo *Clustering* de Enterprise Miner se utiliza especialmente para segmentar bases de datos mediante análisis de conglomerados o *cluster*. Este análisis suele ser útil para segmentar bases de datos. Se busca que los elementos de cada cluster sean similares (homogeneidad dentro) mientras que los elementos en diferentes clusters tienden a ser distintos (heterogeneidad entre). Este método está basado en el uso de la distancia euclídea con método de agrupación de tipo no jerárquico, se basa en el modelo de las *K* medias, lo cual implica que debemos definir *a priori* un rango de número de clusters. Este nombre es debido a que los centros de los clusters son las medias de las observaciones asignadas a cada grupo. El algoritmo tratará de reducir en cada iteración la suma de las distancias al cuadrado en cada grupo hasta que se alcance la convergencia. El nodo incluye un método para detectar outliers para evitar formar grupos de un solo individuo. Es conveniente trabajar con más de 100 observaciones. El nodo clustering se encuentra situado en SAS Enterprise Miner en la categoría *Modify* (Figura 8-1).

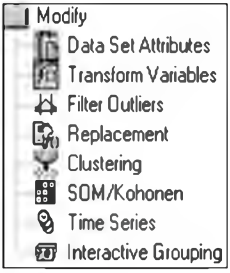


Figura 8-1

Como ejemplo, vamos a utilizar los datos del archivo MUNDO para agrupar, a partir de información de tipo social y económico, a distintos países del mundo en grupos homogéneos. El objetivo es conocer qué países son más o menos parecidos según las variables siguientes: *densidad, urbana, espvidaf, espvidam, alfabet, inc_pob, mortinf, tasa_nat, tasa_mort, tasasida* y *log_pib*. El resto de variables, salvo PA_S, que tiene un papel de etiqueta, serán excluidas del análisis.

Una vez que disponemos ya de nuestros datos de trabajo en formato SAS (archivo *mundo.sas7bdat*) en una determinada librería (librería *Trabajo* que representa el subdirectorio *c:\libros\miningt*), abrimos el proyecto P1 (*File → Open*) y mediante *File → New → Diagram* creamos el diagrama D12. A continuación, el nodo clustering se obtiene mediante el botón *Tools* del navegador de proyecto de Enterprise Miner como subopción de la categoría *Model* o arrastrando el propio nodo sobre la zona de trabajo al lado del nodo *Input Data Source* al que se le ha asignado previamente el conjunto de datos *mundo.sas7bdat* de la librería TRABAJO. A continuación se realizará la unión de ambos (Figura 8-2).

Será necesario declarar las variables *densidad, urbana, espvidaf, espvidam, alfabet, inc_pob, mortinf, tasa_nat, tasa_mort, tasasida* y *log_pib* de tipo *input* y la variable PA_S de tipo *id*, ya que se usará como etiqueta. Esta tarea se realiza haciendo doble clic sobre el nodo *Input Data Source* etiquetado como TRABAJO.MUNDO y eligiendo la pestaña *Variables*. A continuación se hace clic con el botón derecho del ratón sobre cada variable a situar de tipo *input*, se elige *Set Model Role* y a continuación *Input*. Se repite el proceso con las variables a excluir, que se les asignará tipo *rejected* y con la variable PA_S, que se le asignará tipo *id* (Figura 8-3).

 ©Alfaomega-RC Libros

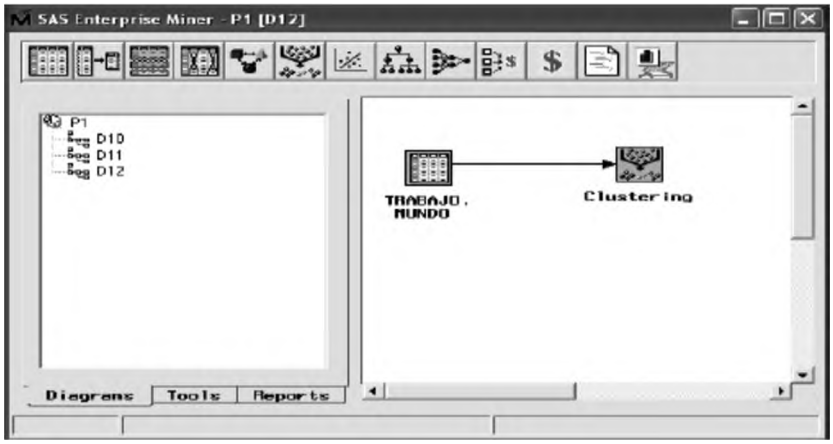
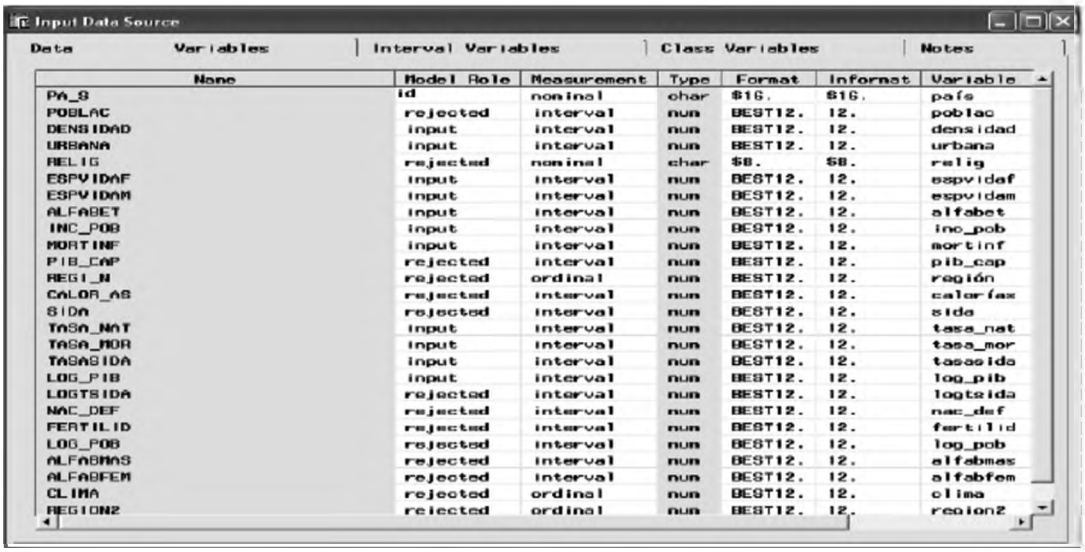


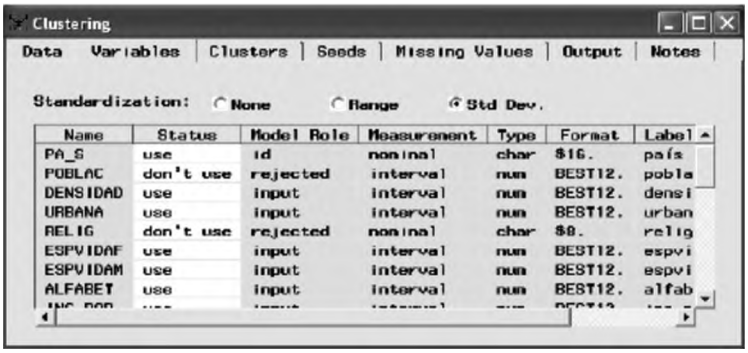
Figura 8-2



Name	Model Role	Measurement	Type	Format	Informat	Variable
PA_S	id	nominal	char	$16.	$16.	país
POBLAC	rejected	interval	num	BEST12.	12.	poblac
DENSIDAD	input	interval	num	BEST12.	12.	densidad
URBANA	input	interval	num	BEST12.	12.	urbana
RELIG	rejected	nominal	char	$8.	$8.	relig
ESPVIDAF	input	interval	num	BEST12.	12.	espvidaf
ESPVIDAM	input	interval	num	BEST12.	12.	espvidam
ALFABET	input	interval	num	BEST12.	12.	alfabet
INC_POB	input	interval	num	BEST12.	12.	inc_pob
MORTINF	input	interval	num	BEST12.	12.	mortinf
PIB_CAP	rejected	interval	num	BEST12.	12.	pib_cap
REGI_N	rejected	ordinal	num	BEST12.	12.	región
CALOR_AS	rejected	interval	num	BEST12.	12.	calorías
SIDA	rejected	interval	num	BEST12.	12.	sida
TASA_NAT	input	interval	num	BEST12.	12.	tasa_nat
TASA_MOR	input	interval	num	BEST12.	12.	tasa_mor
TASASIDA	input	interval	num	BEST12.	12.	tasasida
LOG_PIB	input	interval	num	BEST12.	12.	log_pib
LOGTSIDA	rejected	interval	num	BEST12.	12.	logtsida
NAC_DEF	rejected	interval	num	BEST12.	12.	nac_def
FERTILID	rejected	interval	num	BEST12.	12.	fertilid
LOG_POB	rejected	interval	num	BEST12.	12.	log_pob
ALFABMAS	rejected	interval	num	BEST12.	12.	alfabmas
ALFABFEM	rejected	interval	num	BEST12.	12.	alfabfem
CLIMA	rejected	ordinal	num	BEST12.	12.	clima
REGION2	rejected	ordinal	num	BEST12.	12.	region2

Figura 8-3

Una vez conectados los nodos *Input Data Source* y *Clustering* y definido el papel de cada variable guardaremos la información del nodo de datos y abrimos el nodo clustering haciendo doble clic sobre él en el diagrama. Se obtiene la pantalla de entrada del nodo en cuya pestaña *Variables* elegiremos estandarizar los datos por la desviación típica (Figura 8-4).



Name	Status	Model Role	Measurement	Type	Format	Label
PA_S	use	id	nominal	char	$16.	país
POBLAC	don't use	rejected	interval	num	BEST12.	pobla
DENSIDAD	use	input	interval	num	BEST12.	densi
URBANA	use	input	interval	num	BEST12.	urban
RELIG	don't use	rejected	nominal	char	$8.	relig
ESPVIDAF	use	input	interval	num	BEST12.	espvi
ESPVIDAM	use	input	interval	num	BEST12.	espvi
ALFABET	use	input	interval	num	BEST12.	alfab

Figura 8-4

A continuación, en la pestaña *Clusters*, podemos indicar que queremos una solución que incluya 6 grupos de países. Este parámetro podemos definirlo en la opción *Number of Clusters* (Figura 8-5). En la pestaña *Mising Values* elegimos como método de imputación de valores desaparecidos el de la media del cluster o grupo más cercano (Figura 8-6). En la ventana *Output* (pestaña *Print*) pediremos que en los resultados nos muestre las estadísticas de los *Cluster* y el grupo de países que conforma cada grupo (Figura 8-7).

Figura 8-5

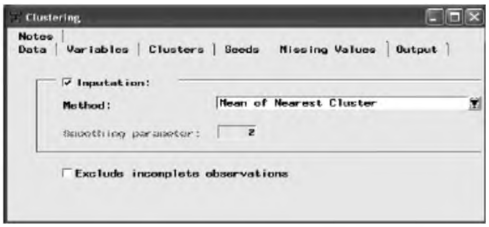


Figura 8-6

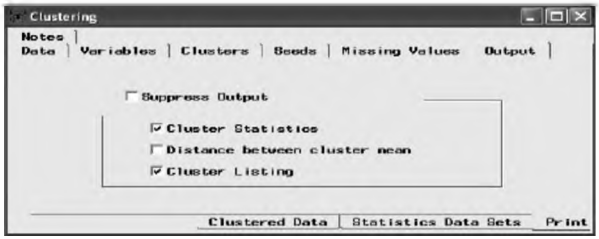


Figura 8-7

A continuación cerramos el nodo (botón) y después lo ejecutaremos haciendo clic sobre él en el diagrama con el botón derecho del ratón y eligiendo la opción *Run* del menú emergente resultante (Figura 8-8). Finalizada la ejecución el sistema pregunta si queremos ver los resultados (Figura 8-9). Al aceptar se abre el *visor de resultados* cuya pestaña *Partition* (Figura 8-10) muestra una idea gráfica de los clusters.

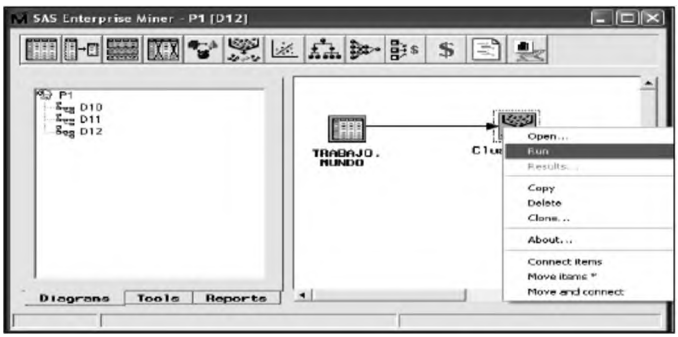


Figura 8-8

©Alfaomega-RC Libros

Figura 8-9

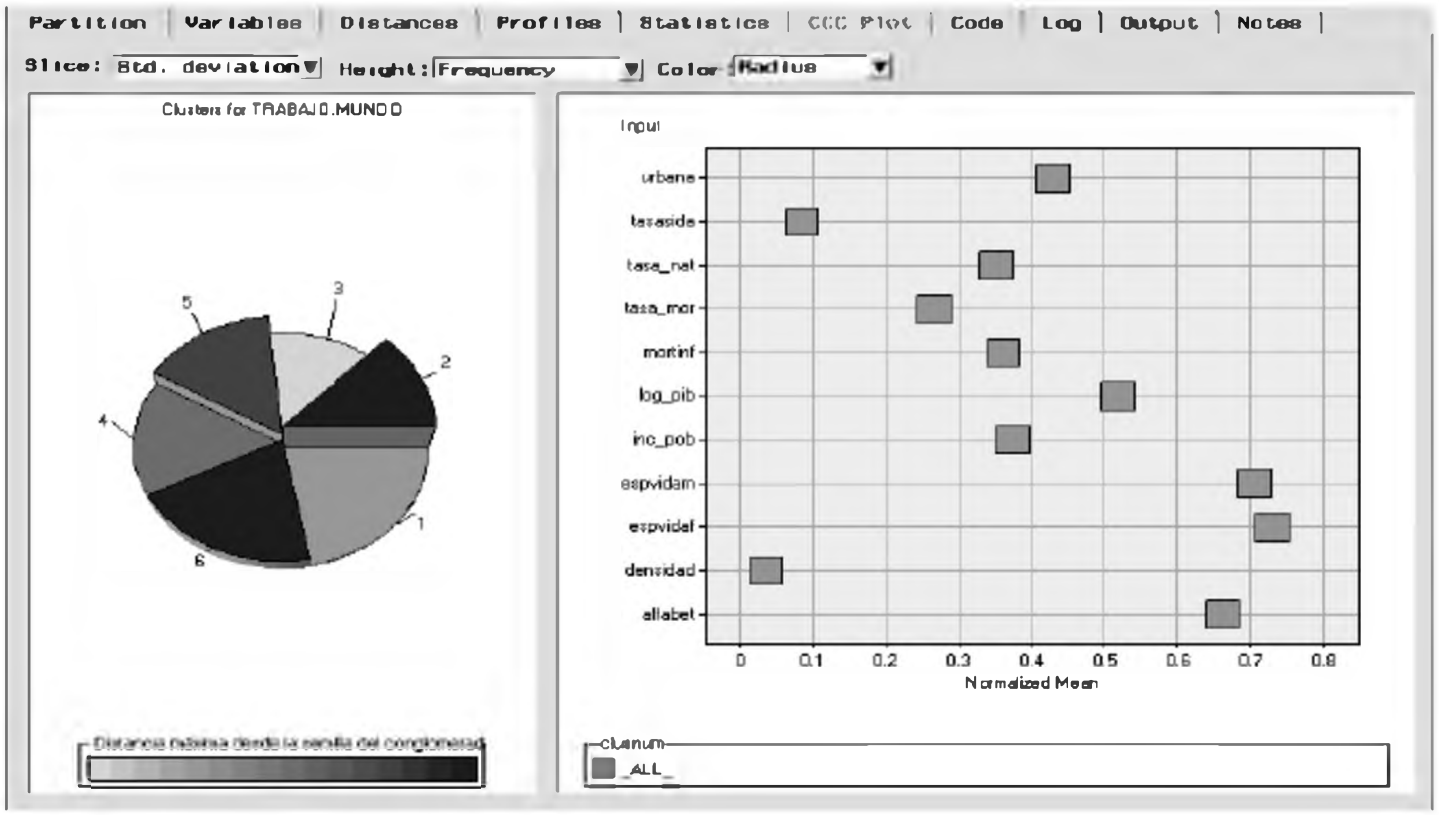
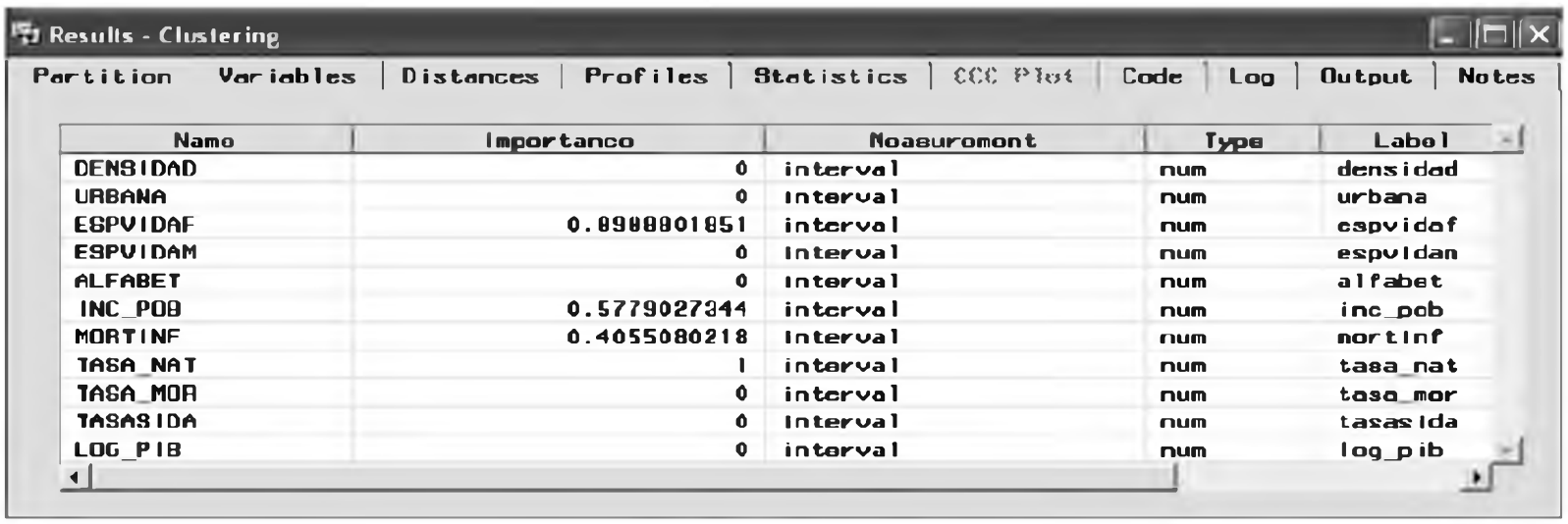


Figura 8-10

La pestaña *Variables* muestra la importancia relativa de cada una de las variables utilizadas para hacer los grupos (Figura 8-11).



Name	Importance	Measurement	Type	Label
DENSIDAD	0	interval	num	densidad
URBANA	0	interval	num	urbana
ESPVIDAF	0.8988801851	interval	num	espvidaf
ESPVIDAM	0	interval	num	espvidam
ALFABET	0	interval	num	alfabet
INC_POB	0.5779027344	interval	num	inc_pob
MORTINF	0.4055080218	interval	num	mortinf
TASA_NAT	1	interval	num	tasa_nat
TASA_MOR	0	interval	num	tasa_mor
TASASIDA	0	interval	num	tasasida
LOG_PIB	0	interval	num	log_pib

Figura 8-11

La pestaña *Distances* muestra que los grupos están relativamente cerca. Solo el grupo 3 parece ser mucho más distinto que los demás (Figura 8-12).

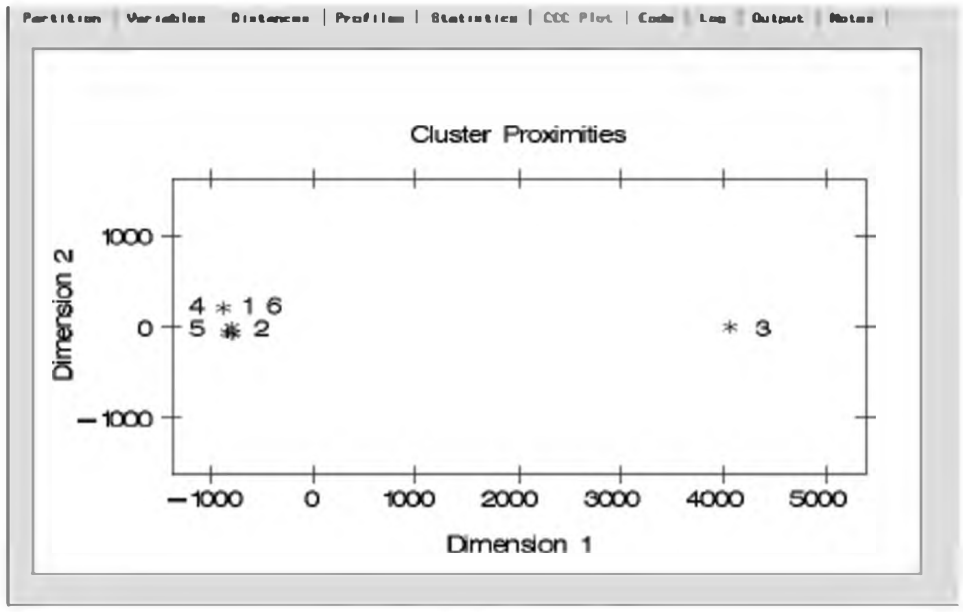


Figura 8-12

La pestaña *Statistics* muestra el número de individuos en cada cluster, la varianza intra grupo, la distancia máxima de un país al centro del cluster y el cluster más cercano, etc. (Figura 8-13).

CLUSTER	Frecuencia del conglomerado	Desviación estándar de la raíz cuadrada	Distancia máxima desde la semilla del co	Conglomerado más cercano	Distancia hast
1	2	0.6847397476	1.6059547066	6	5.1022966991
2	45	0.4017126615	2.9802133874	5	2.8727692[illegible]
3	2	0.4225281189	0.9939162738	2	7.3501779204
4	10	0.5073636452	2.0163182409	5	2.0450136545
5	30	0.4591475533	2.4785674593	4	2.0450136545
6	20	0.6364456464	2.8898787732	5	4.3315327216

Figura 8-13

La pestaña *Code* presenta el código SAS del procedimiento (Figura 8-14).

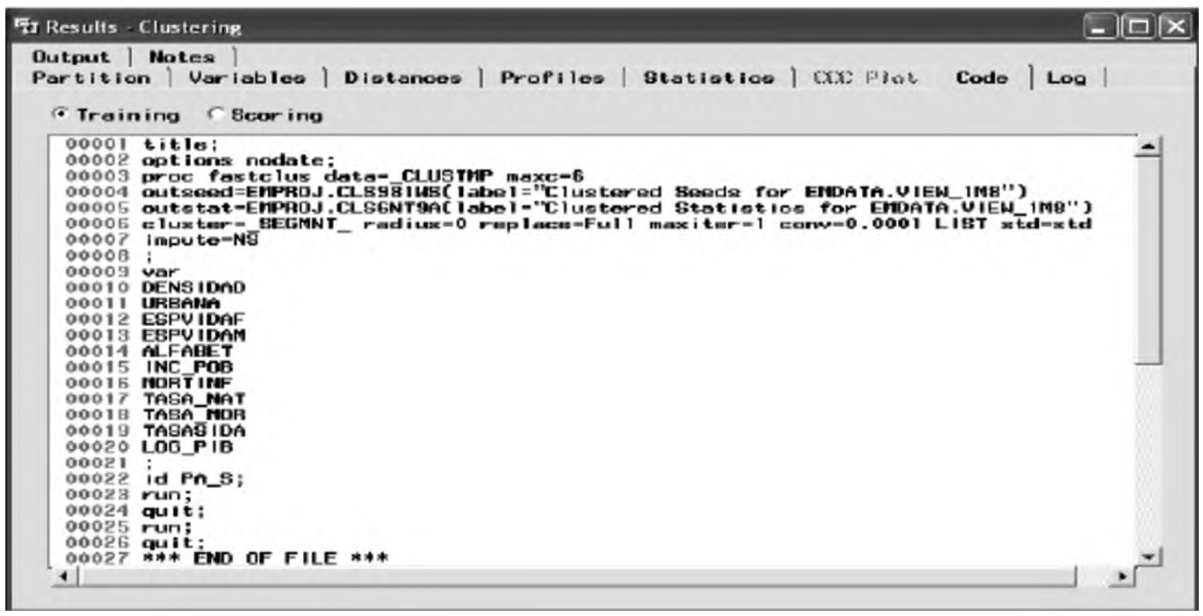


Figura 8-14

La pestaña *Output* lista las medias en las variables utilizadas para cada cluster así como el grupo al que pertenece cada país (Figura 8-15).

©Alfaomega-RC Libros

Results - Clustering

Partition | Variables | Distances | Profiles | Statistics | CCC Plot | Code | Log | Output | Notes

Procedimiento FASTCLUS
Replace=FULL Radius=0 Maxclusters=6 Impute Maxiter=1

Semillas iniciales

Conglomerado	densidad	urbana	espvidaf	espvidam	alfabet	inc_pob
1	0.016279269	1.735295776	4.256616112	4.745087307	3.190114461	2.338086049
2	0.350744677	3.677174383	7.567317533	7.980374108	4.326319612	0.167006146
3	8.130764802	3.883757213	7.567317533	8.088217001	3.364915254	-0.075152766
4	0.143553729	3.966390346	7.378134594	7.872531215	3.190114461	4.375561035
5	0.003551845	1.032914152	6.243036964	6.470573601	3.146414263	2.254582976
6	1.183948278	0.661065058	5.013347865	5.715673348	1.529506934	2.004073757

Semillas iniciales

Conglomerado	mortinf	tasa_nat	tasa_mor	tasasida	log_pib
1	2.232189278	3.721412662	4.232076624	6.606964025	4.448241548
2	0.189079478	1.051703578	2.586269048	0.312481536	6.779097599
3	0.152314024	1.051703578	1.410692208	0.034514151	6.718066018
4	0.328262982	2.265207707	0.470230736	0.011233562	6.182768797
5	1.032058816	2.588808808	1.880922944	2.043705913	5.527974311
6	2.783670088	2.831509634	2.586269048	0.000017342	3.717978558

Listado de conglomerados

Observación	país	Cluster	Distancia del valor germinal
1	Acerbaján	5	1.6480
2	Afganistán	6	2.8080
3	Alemania	2	1.0609
4	Arabia Saudí	4	0.8382

Figura 8-15

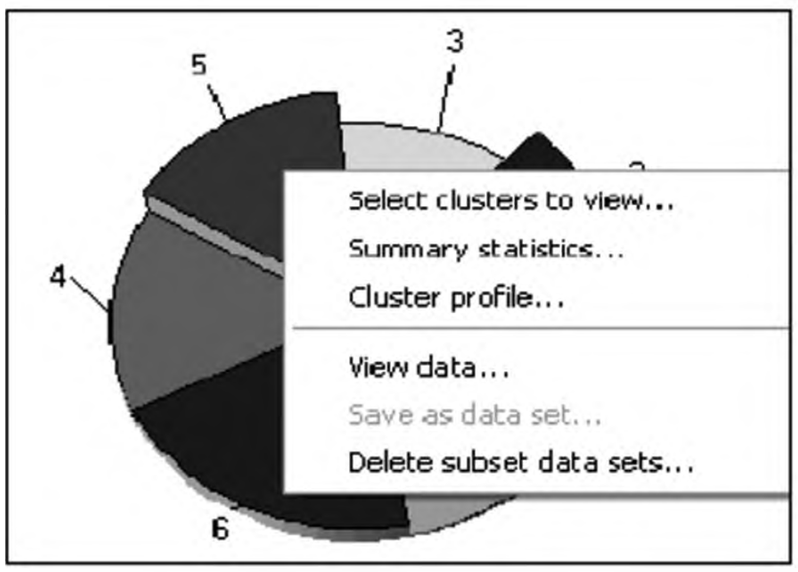


Figura 8-16

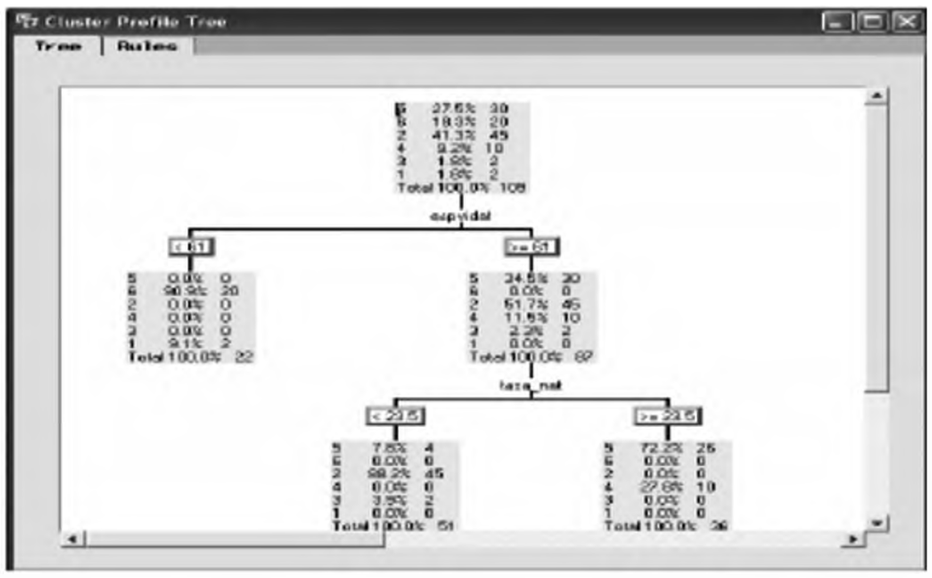


Figura 8-17

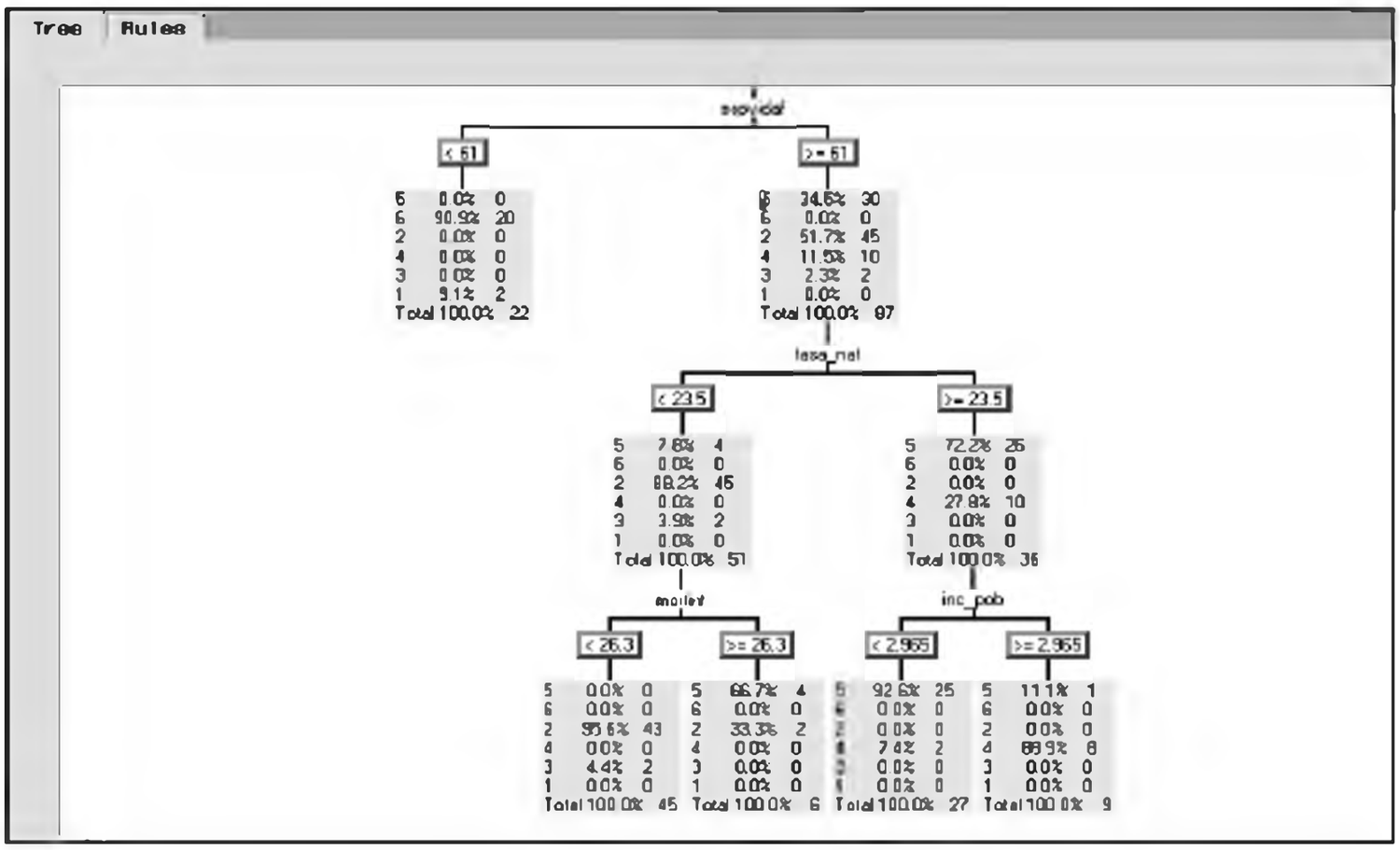


Figura 8-18

Si en la Figura 8-16 de la pestaña *Partition* hacemos clic con el botón derecho del ratón sobre el gráfico de sectores, obtenemos el menú emergente de la Figura 8-16. La opción *Cluster profile* de este menú muestra el árbol de decisión que permite asignar los casos a un cluster determinado (Figuras 8-17 y 8-18). Junto a este gráfico, seleccionando la pestaña *Rules* podemos ver las reglas lógicas que caracterizan el árbol.

ANÁLISIS CLUSTER CON REDES NEURONALES: NODO SOM/KOHONEN

El nodo *SOM/Kohonen,* situado en el grupo *Modify* del menú de Miner (Figura 8-19), es un método para llevar a cabo análisis *cluster* con redes neuronales artificiales y aprendizaje no supervisado, esto es, no se especificará una variable de resultados. Este método permite también realizar *Kohonen Vector Quantization* (VQ). El término *SOM* procede del término inglés *Self-Organizing Map* mientras que *Kohonen* es el apellido de Teuvo Kohonen, analista que propuso este tipo de redes neuronales auto organizadas. En esta ocasión se suministran los inputs de entrada, que pueden ser de tipo binario, nominal, ordinal o continuo, pero la red no recibe ninguna información que le indique cuál es la salida. La red, que solo consta de dos capas y modifica los pesos para poder asociar a la misma unidad de salida los vectores de entrada más similares. Suele decirse así que estas redes son capaces de auto-organizarse. El método Kohonen VQ está orientado a la clusterización mientras que el método SOM puede ser utilizado como método de reducción de la dimensión.

Como ejemplo, vamos a utilizar los datos del archivo *MUNDO* para agrupar, a partir de información de tipo social y económico, a distintos países del mundo en grupos homogéneos. El objetivo es conocer qué países son más o menos parecidos según las variables siguientes: *densidad, urbana, espvidaf, espvidam, alfabet, inc_pob, mortinf, tasa_nat, tasa_mort, tasasida* y *log_pib*. El resto de variables, salvo *PA_S*, que tiene un papel de etiqueta y serán excluidas del análisis.

Una vez que disponemos ya de nuestros datos de trabajo en formato SAS (archivo *mundo.sas7bdat*) en una determinada librería (librería *Trabajo* que representa el subdirectorio *c:\libros\miningt*), abrimos el proyecto P1 (*File* $\rightarrow$ *Open*) y mediante *File* $\rightarrow$ *New* $\rightarrow$ *Diagram* creamos el diagrama D17. A continuación, el nodo *SOM/Kohonen* se obtiene mediante el botón *Tools* del navegador de proyecto de Enterprise Miner como subopción de la categoría *Modify* o arrastrando el propio nodo sobre la zona de trabajo al lado del nodo *Input Data Source* al que se le ha asignado previamente el conjunto de datos *mundo.sas7bdat* de la librería TRABAJO. A continuación se realizará la unión de ambos (Figura 8-19).

 ©Alfaomega-RC Libros

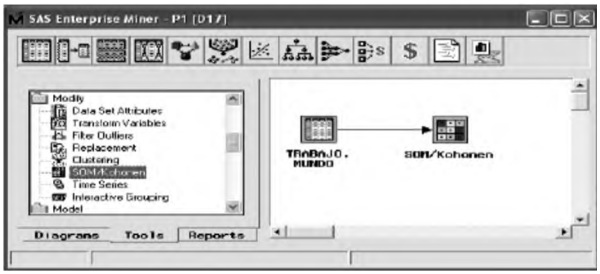
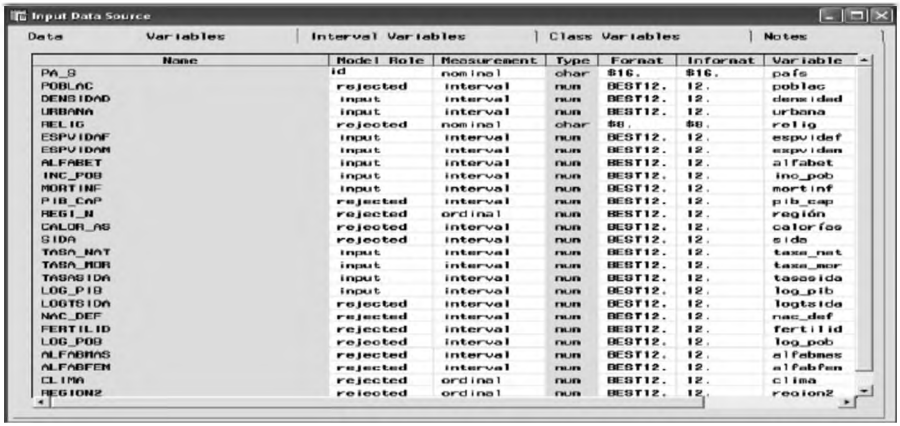


Figura 8-19

Será necesario declarar las variables *densidad, urbana, espvidaf, espvidam, alfabet, inc_pob, mortinf, tasa_nat, tasa_mort, tasasida* y *log_pib* de tipo *input* y la variable *PA_S* de tipo *id*, ya que se usará como etiqueta. Esta tarea se realiza haciendo doble clic sobre el nodo *Input Data Source* etiquetado como TRABAJO.MUNDO y eligiendo la pestaña *Variables*. A continuación pulse con el botón derecho del ratón sobre cada variable a situar de tipo *input*, se elige *Set Model Role* y pulse *Input*. Se repite el proceso con las variables a excluir, que se les asignará tipo *rejected* y con la variable *PA_S*, que se le asignará tipo *id* (Figura 8-20). Una vez conectados los nodos *Input Data Source* y *SOM/Kohonen* y definido el papel de cada variable guardaremos la información del nodo de datos y abrimos el nodo *SOM/Kohonen* haciendo doble clic sobre él en el diagrama. Se obtiene la pantalla de entrada del nodo en cuya pestaña *Variables* elegiremos las variables que serán usadas y si estas serán o no estandarizadas por rango o desviación típica (Figura 8-21).

La ventana *Cluster* (Figura 8-22) contiene los campos *Variable name* (nombre para el identificador del cluster), *Variable label* (etiqueta del *cluster*) y *Role* (papel que se asigna a la variable que está siendo usada para formar los *clusters* que por defecto es de grupo o *group*). En la ventana *General* se pueden seleccionar el método, el tamaño del mapa y el número de *clusters* (Figura 8-23).



Name	Model Role	Measurement	Type	Format	Informat	Variable
PA_S	id	nominal	char	$16.	$16.	país
POBLAC	rejected	interval	num	BEST12.	12.	poblac
DENSIDAD	input	interval	num	BEST12.	12.	densidad
URBANA	input	interval	num	BEST12.	12.	urbana
RELIG	rejected	nominal	char	$8.	$8.	relig
ESPVIDAF	input	interval	num	BEST12.	12.	espvidaf
ESPVIDAM	input	interval	num	BEST12.	12.	espvidam
ALFABET	input	interval	num	BEST12.	12.	alfabet
INC_POB	input	interval	num	BEST12.	12.	inc_pob
MORTINF	input	interval	num	BEST12.	12.	mortinf
PIB_CAP	rejected	interval	num	BEST12.	12.	pib_cap
REGI_N	rejected	ordinal	num	BEST12.	12.	región
CALOR_AS	rejected	interval	num	BEST12.	12.	calorías
SIDA	rejected	interval	num	BEST12.	12.	sida
TASA_NAT	input	interval	num	BEST12.	12.	tasa_nat
TASA_MOR	input	interval	num	BEST12.	12.	tasa_mor
TASASIDA	input	interval	num	BEST12.	12.	tasasida
LOG_PIB	input	interval	num	BEST12.	12.	log_pib
LOGTSIDA	rejected	interval	num	BEST12.	12.	logtsida
NAC_DEF	rejected	interval	num	BEST12.	12.	nac_def
FERTILID	rejected	interval	num	BEST12.	12.	fertilid
LOG_POB	rejected	interval	num	BEST12.	12.	log_pob
ALFABMAS	rejected	interval	num	BEST12.	12.	alfabmas
ALFABFEM	rejected	interval	num	BEST12.	12.	alfabfem
CLIMA	rejected	ordinal	num	BEST12.	12.	clima
REGION2	rejected	ordinal	num	BEST12.	12.	region2

Figura 8-20

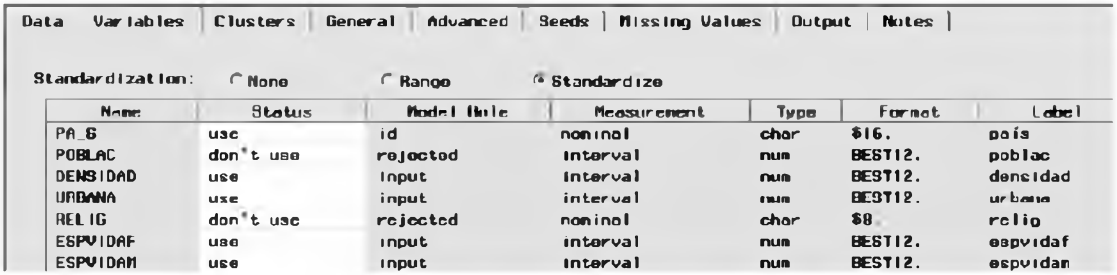


Figura 8-21

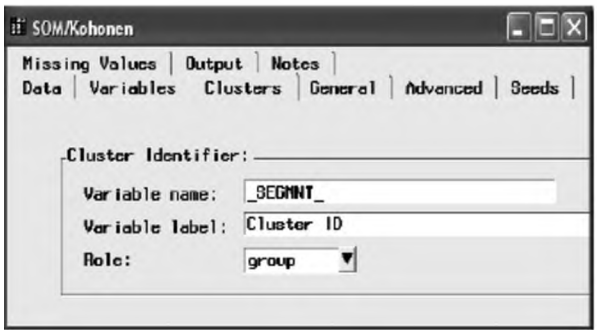


Figura 8-22

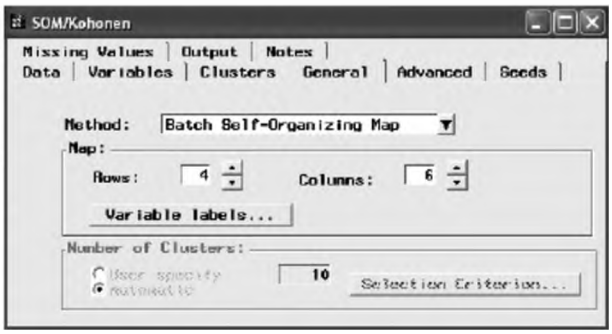


Figura 8-23

A continuación se explican los campos de la ventana *General* (Figura 8-23).

Método: seleccione una de las siguientes opciones:

- ***Batch Self-Organizing Map:*** las opciones más importantes son la forma del mapa y el tamaño del vecino final. Normalmente son preferibles mapas grandes si bien ello aumenta el tiempo de entrenamiento.

- ***Kohonen Self-Organizing Map:*** las opciones más importantes son la forma del mapa y el coeficiente de aprendizaje. Es importante empezar con un coeficiente de aprendizaje alto como 0,9.

- ***Kohonen Vector Quantization:*** las opciones más importantes son el número de *clusters* y el coeficiente de aprendizaje. La elección del número de *clusters* óptimo se obtiene por ensayo-error.

Mapa: selecciona el número de filas y columnas que por defecto es de 4 y 6, respectivamente. El número óptimo se encuentra mediante un proceso ensayo-error. Si el mapa es muy pequeño seguramente los grupos no reflejarán las no linealidades que contengan los datos. Si el mapa es muy grande el tiempo de análisis será elevado y la formación de *clusters* vacíos tenderá a dificultar la interpretación de los resultados. Al cambiar este tamaño también variaremos proporcionalmente el tamaño del vecino

©Alfaomega-RC Libros

(*Neighborhood Options* en la pestaña *Advanced*). Podemos también renombrar las variables de fila y columna del mapa en la opción *Variable labels*.

Número de clusters: se activa al seleccionar el método de entrenamiento Kohonen VQ. Por defecto se determina de forma automática si bien puede ser especificado por el usuario. Podemos también como en el nodo *Clustering* variar el método de *clustering*, el *clustering cubic criterion cutoff* y el número mínimo y máximo de *clusters* en la opción *Selection Criterion*.

En la ventana *Advanced* se pueden seleccionar las opciones avanzadas en las subventanas: *Batch SOM Training, Neighborhood Options* y *Kohonen Training* (Figura 8-24).

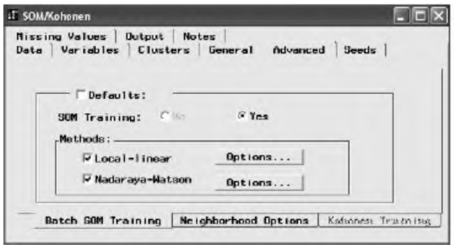


Figura 8-24

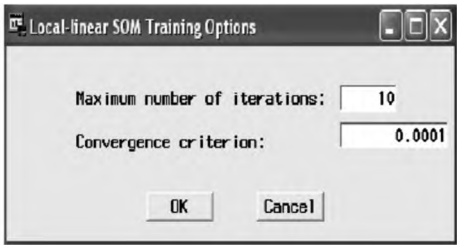


Figura 8-25

A continuación se explican las subventanas de la ventana *Advanced* (Figura 8-24).

Batch SOM Training: por defecto las opciones se determinan en función del tipo de red elegida en la pestaña *General*. Para *batch SOM*: se activa *SOM Training, Local-linear* y *Nadaraya-Watson*. Para *Kohonen SOM* se desactivan las opciones anteriores. Si *Local-linear* y *Nadaraya-Watson* están activadas al mismo tiempo, primero se realiza el entrenamiento *Nadaraya-Watson* y después *Local-linear*. Los criterios de convergencia en ambos, bien número de iteraciones, bien un criterio establecido, se definen en la opción *Options* (Figura 8-25).

Neighborhood Options: al seleccionar uno de los métodos SOM en la ventana *General* podremos establecer distintas opciones de vecindad deseleccionando *Defaults* (Figura 8-26).

- ***Kernel shape:*** por defecto es *Epanechnikov* (1). Otras formas son uniforme (0), *biweight* (2), *triweight* (3) y otras (valor *k* positivo definido por el analista).

- ***Kernel metric:*** por defecto es *Max* (0). Otras son *Manhattan* o *cityblock* (1), *euclídea* (2) y otros valores positivos especificados por el analista.

- ***Size:*** el valor de vecindad debe ser mayor que o igual a cero. Seleccionando *Options* podemos establecer el tamaño inicial, final, número de pasos hasta alcanzar el tamaño final y el número de iteraciones hasta alcanzar el tamaño final (Figura 8-27).

- ***Number of steps before reset:*** número de pasos antes de terminar.

Sea *Row(j)* el número de fila del *cluster j*-ésimo, *Col(j)* el número de columna del *cluster j*-ésimo, *size* el tamaño de vecindad, *k* la forma *kernel* y *p* la métrica *kernel*. Entonces la función *kernel* se define como:

$$K(j,n) = \left\{ 1 - \frac{\left[|Row(j) - Row(n)|^p + |Col(j) - Col(n)|^p \right]^{2/p}}{size^2} \right\}^{2k}$$

Figura 8-26

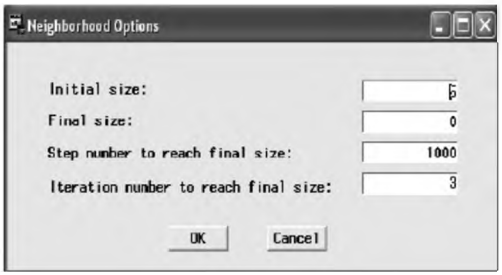


Figura 8-27

Así para un *kernel* uniforme $K(j,n)=1$ cuando la distancia entre dos semillas iguala el tamaño de vecindad (*size*) mientras que para otros *kernels* $K(j,n)=0$ cumpliéndose de nuevo que la distancia entre dos semillas iguala el tamaño de vecindad (*size*).

Kohonen Training: al seleccionar *Kohonen SOM* o *Kohonen VQ* en la ventana *General* podremos establecer distintas opciones avanzadas para el entrenamiento de Kohonen en la subventana *Kohonen Training* de la ventana *Advanced* (Figura 8-28).

- ***Learning rate:*** se utiliza el entrenamiento que varía el coeficiente de aprendizaje. Este se inicializa en 0,9 para Kohonen SOM y 0,5 para VQ. Se reduce linealmente en 0,02 durante los primeros 1000 pasos del entrenamiento. El coeficiente de aprendizaje, que debe variar entre 0 y 1, puede ser redefinido seleccionando *Options*...

©Alfaomega-RC Libros

- ***Maximum number of steps:*** por defecto es de 500 veces el número de clusters. Un paso es el proceso llevado a cabo en cada elemento.
- ***Maximum number of iterations:*** por defecto es 100. Una iteración es el proceso llevado a cabo sobre todos los datos de entrenamiento.
- ***Convergence criterion***: por defecto es 0.0001.

El entrenamiento finaliza cuando se satisface alguno de los criterios de convergencia.

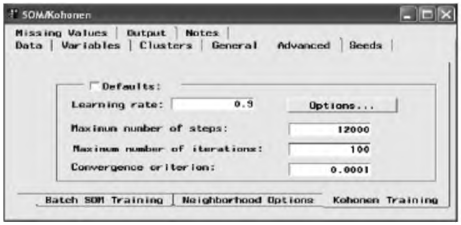


Figura 8-28

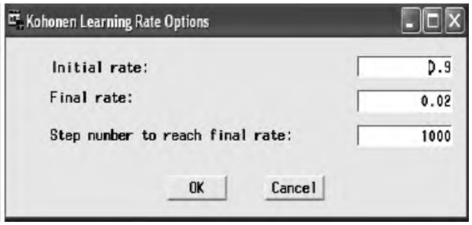


Figura 8-29

La ventana *Seeds* en el nodo SOM/Kohonen permite especificar la forma de inicializar las semillas de los *clusters* (Figura 8-64). Por defecto dependerán del tipo de red. A *Batch SOM* le corresponde *Principal Component*. A *Kohonen SOM* y a *Kohonen VQ* le corresponde *random*. Para otros métodos las semillas iniciales son casos de entrenamiento sin datos perdidos separados por una distancia mínima especificada en la opción *Minimum distance between cluster seeds* (*radius*).

La ventana *Missing Values* en el nodo *SOM/Kohonen* permite establecer cómo tratar los casos con datos perdidos (Figura 8-31). Las observaciones que presentan datos perdidos en todas las variables son excluidas del análisis. Los campos de esta ventana se explican a continuación:

Imputation method: podemos elegir entre los siguientes métodos de imputación.

- ***Seed of Nearest Cluster****:* semilla del *cluster* más cercano.
- ***Ignore****:* ignorar.

Processing of Missing Values during Training*:* existen 7 métodos dependiendo alguno de ellos del tipo de variable.

- ***Ignore****:* ignorar el procesado.
- ***Category****:* asignar a una categoría.

- ***Mean**:* asignar la media.
- ***Median**:* asignar la mediana.
- ***Midrange**:* asignar el valor promedio del rango.
- ***Mode**:* asignar la moda.
- ***Omit**:* omitir el elemento.

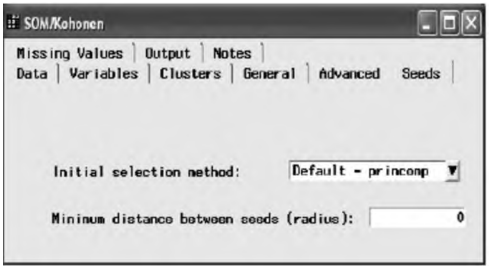


Figura 8-30

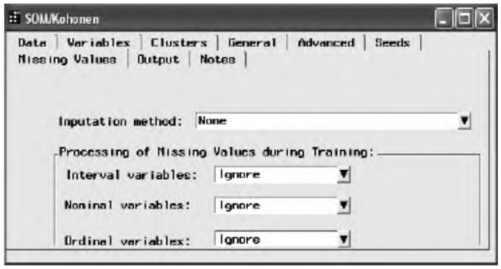


Figura 8-31

La ventana de *Output* en el nodo *SOM/Kohonen* consta de dos subventanas:

Clustered Data: muestra las librerías y los ficheros de resultados para los datos de entrenamiento, validación y test (Figura 8-32). Es importante destacar que estos resultados se guardan en la misma librería que el proyecto en el que está contenido el diagrama. Los datos de resultados del *cluster* contienen los datos originales, una variable de segmentación y una variable de la distancia a la que se asigna el papel *rejected*. Filas (*Row*), Columnas (*Column*) y variables Filas: Columnas (coordenadas del mapa) son guardadas en el fichero *output* cuando un método SOM es usado en el entrenamiento. Si los datos perdidos han sido imputados entonces se añadirá una variable denominada *_IMPUTE_* al fichero de resultados.

Statistics Data Sets: en esta ventana se muestran las bases de datos que contienen las estadísticas del análisis *cluster* y de las semillas (Figura 8-33). Esta última es especialmente útil para clasificar otros datos usando el comando de SAS FASTCLUS.

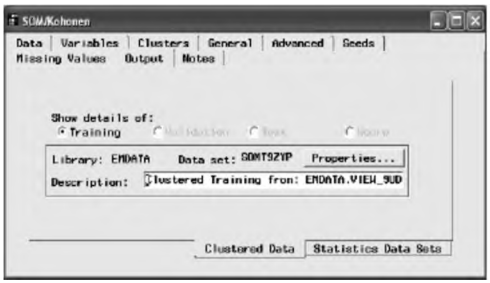


Figura 8-32

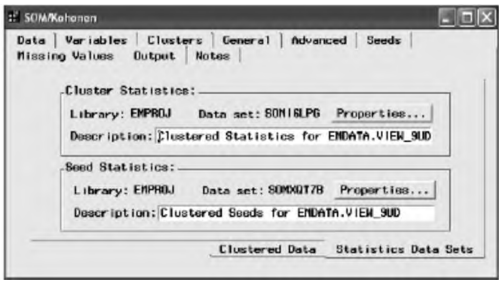


Figura 8-33

©Alfaomega-RC Libros

Una vez seleccionadas todas las opciones estaremos en disposición de ejecutar el nodo. Para nuestro ejemplo hemos ido tomando las opciones por defecto salvo en la pestaña *Cluster* en la que hemos elegido el método *Kohonen Vector Quantization* y hemos fijado 6 *clusters* (Figura 8-34). Tras ejecutar el análisis se abre el *Visor de resultados*.

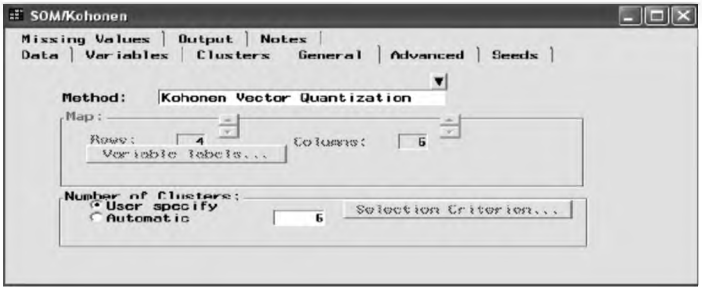


Figura 8-34

Ejecutado el nodo, vemos que los resultados del visor de resultados son equivalentes a los del nodo de *clustering* ya visto en el capítulo de análisis cluster. Por ejemplo, la pestaña *Map* (Figura 8-35) presenta la misma salida.

Si hacemos clic con el botón secundario del ratón sobre cualquier punto del gráfico de sectores de la ventana *Map* se obtiene el menú emergente de la Figura 8-36, cuya opción *View Data* nos lleva al cuadro de datos de la Figura 8-37 que presenta todas las variables del archivo de datos y el *cluster* al que pertenecen (columna CLUSTER_ID).

La opción *Cluster Profile* del menú emergente anterior, nos lleva a un gráfico que presenta el perfil de formación de los *clusters* de un modo arborescente (Figura 8-38).

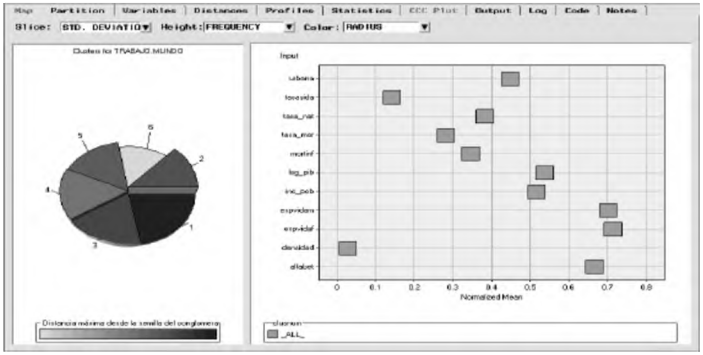


Figura 8-35

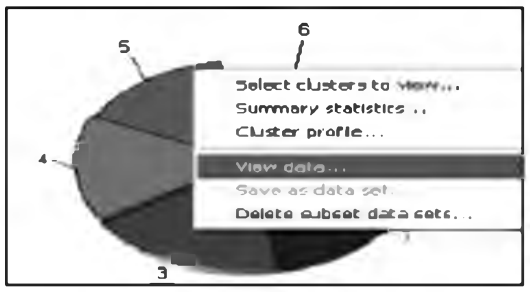


Figura 8-36

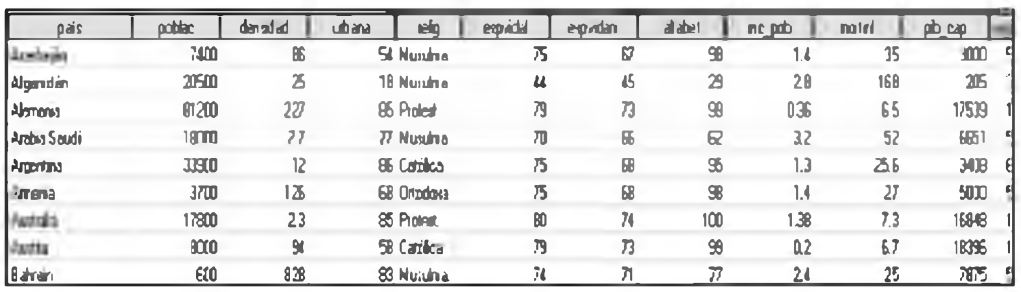


Figura 8-37

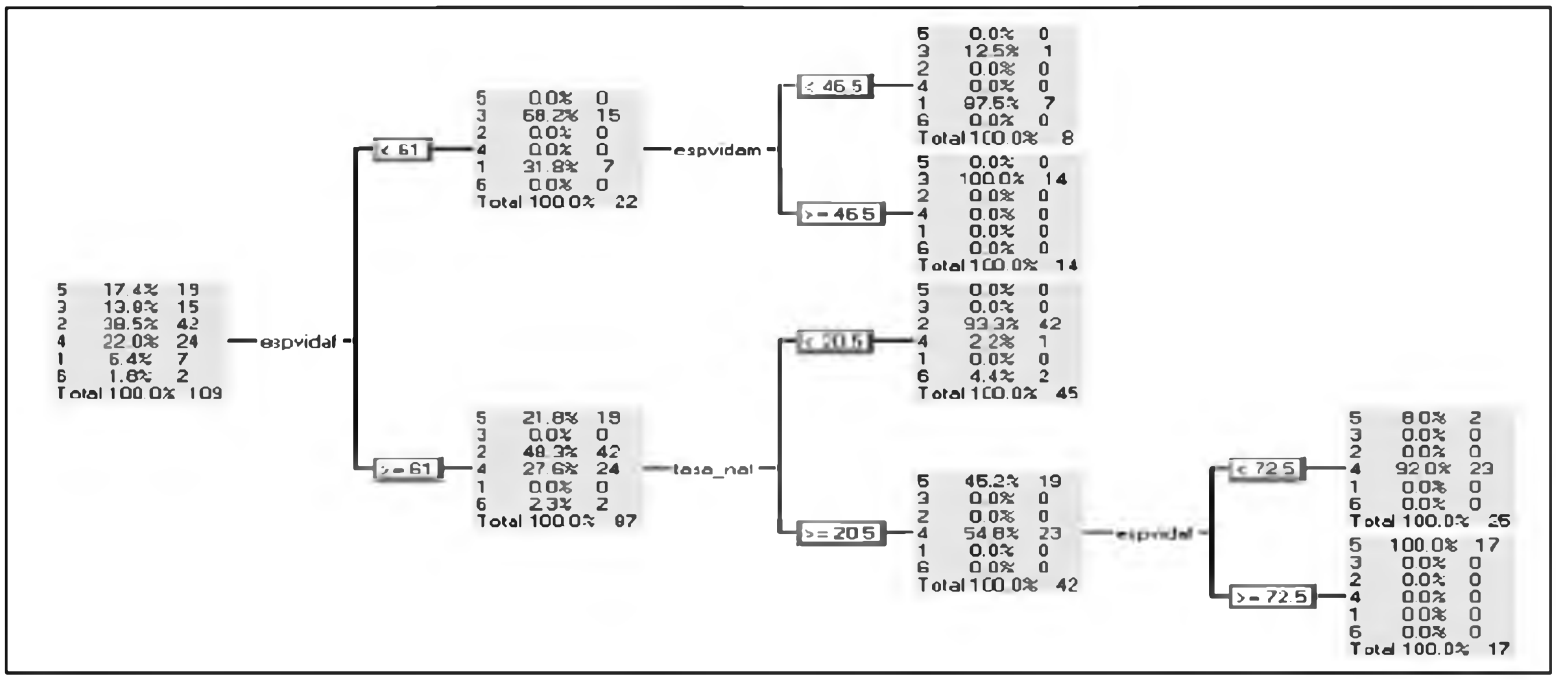


Figura 8-38

EL PROCESO DE CREACIÓN DE ÁRBOLES DE DECISIÓN CON SAS ENTERPRISE MINER

Con los datos del fichero en formato ASCII de nombre *tenis.txt* que contiene información acerca de los días que se ha podido jugar al tenis en función de diversos aspectos meteorológicos, se trata de construir un modelo basado en árboles de decisión que permita predecir si a día de hoy es posible jugar al tenis. El problema es el mismo resuelto ya en el caso de SPSS Modeler. Los datos se presentan en la Figura 8-39.

Figura 8-39

©Alfaomega-RC Libros

Nodo Input Data Source para leer ficheros de datos y enlazarlos con SAS Enterprise Miner

La primera tarea en un proceso de minería de datos es leer el conjunto de datos de partida. Esta tarea la realiza *SAS Enterprise Miner* a través del nodo *Input Data Source* (nodo de datos). El nodo de datos es un nodo fundamental, ya que permite leer los datos del análisis y definir sus atributos. El nodo solo lee los datos en formato SAS, es por ello que si nuestros datos están en otro formato, en nuestro caso en ASCII delimitado por columnas o tabuladores, deberemos importarlos a alguna de las librerías activas. Si es necesario, se crea una nueva librería.

Para crear una librería pulsaremos el icono y en los campos de la Figura 8-40 nombraremos la nueva librería (*Trabajo*) y activaremos el cuadro *Asignar al inicio* para que esté disponible cada vez que abramos SAS.

Para importar el archivo ASCII *tenis.txt* a SAS, en el menú principal de SAS seleccionaremos *Archivo → Importar Datos* eligiendo *Fichero delimitado por Tab (.txt)* según se indica en la Figura 8-41. Se pulsa *Siguiente* y se elige el camino en que se encuentra el fichero a importar (Figura 8-42). Se pulsa *Siguiente* otra vez y se elige la librería en que se alojará el nuevo fichero SAS importado y su nombre (Figura 8-43). Al pulsar *Finalizar* se obtiene ya el fichero importado en formato SAS *tenis.sas7bdat* que se aloja en la librería *Trabajo* (*C:\libros\miningt\datos*).

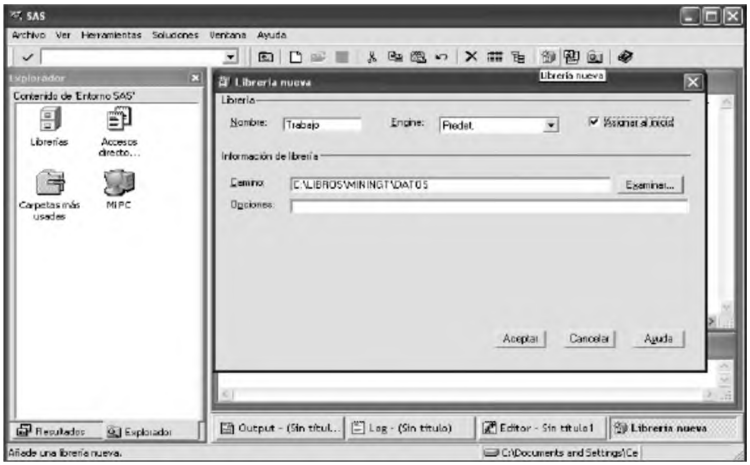


Figura 8-40

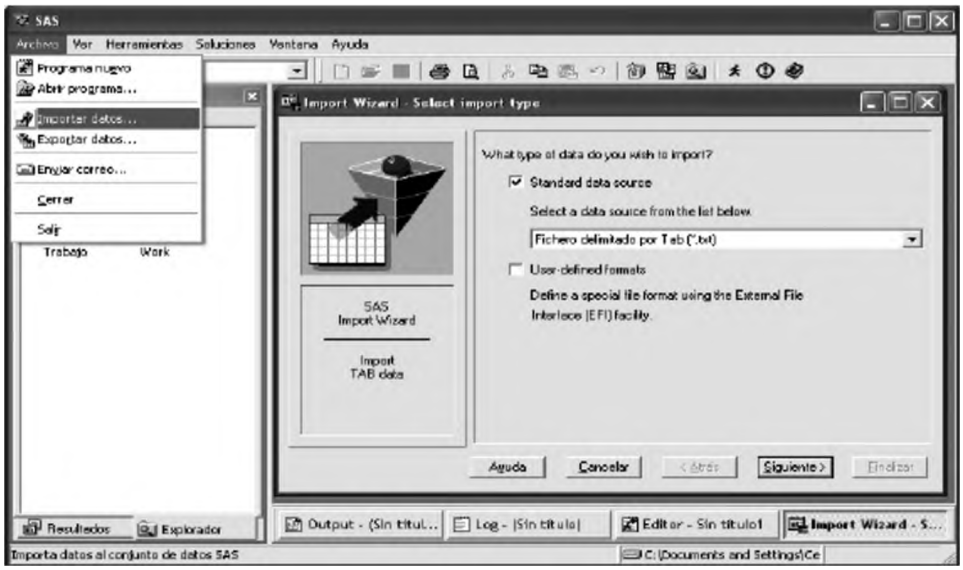


Figura 8-41

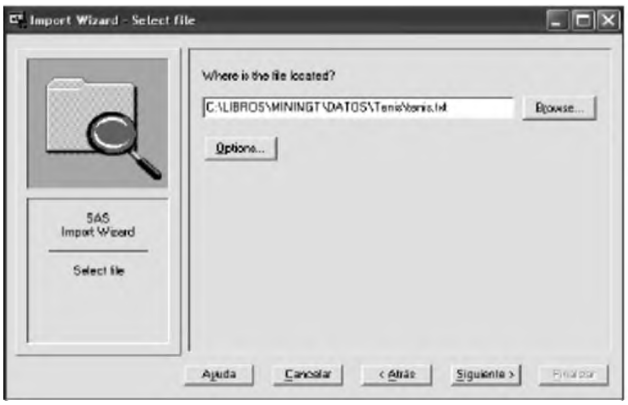


Figura 8-42

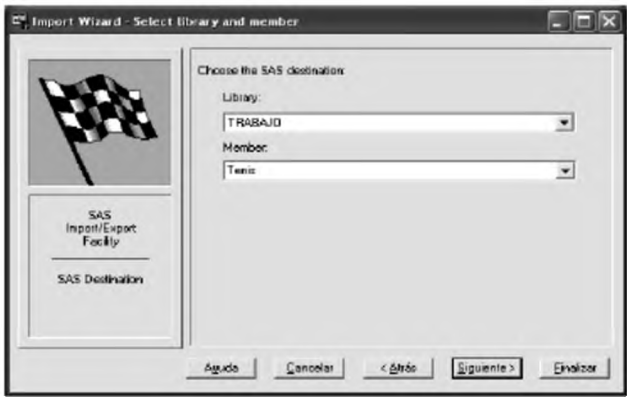


Figura 8-43

En el *log* de SAS se obtiene un mensaje informando de la creación del nuevo fichero (Figura 8-44). Si queremos ver su contenido basta con hacer clic en la librería *Trabajo* y hacer doble clic sobre el fichero *Tenis*. Se abre una ventana con el contenido del fichero SAS (Figura 8-45). De forma similar se importan datos desde Excel, Access, dBase, Lotus, etc.

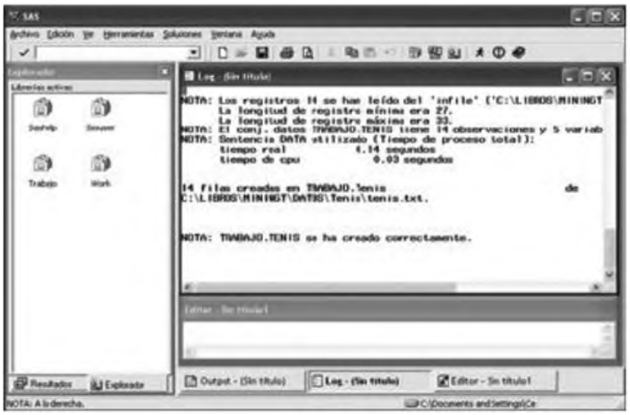


Figura 8-44

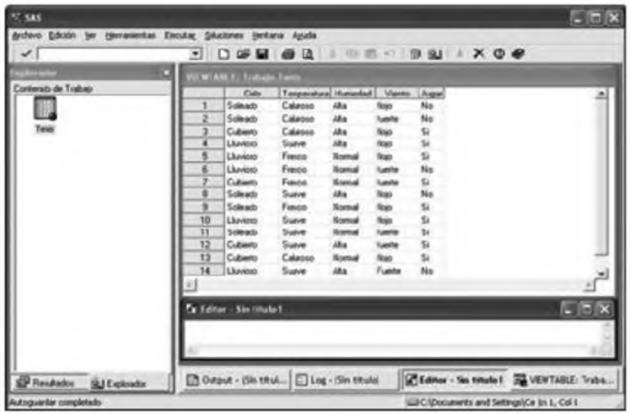


Figura 8-45

©Alfaomega-RC Libros

Una vez que disponemos ya de nuestros datos de trabajo en formato SAS en una determinada librería, ejecutamos *SAS Enterprise Miner* tecleando *miner* en la caja de comandos de SAS y pulsando *Enter* o mediante *Soluciones* → *Análisis* → *Enterprise Miner*. Al hacer clic en el botón *Tools* del navegador de proyecto se obtiene el listado de herramientas de *Enterprise Miner* entre las que se observa el nodo *Input Data Source* (Figura 8-46) como subopción de la categoría *Sample*. El nodo se inserta en la zona de trabajo arrastrándolo hasta ella desde el listado de herramientas o desde el icono de la barra de herramientas (Figura 8-47).

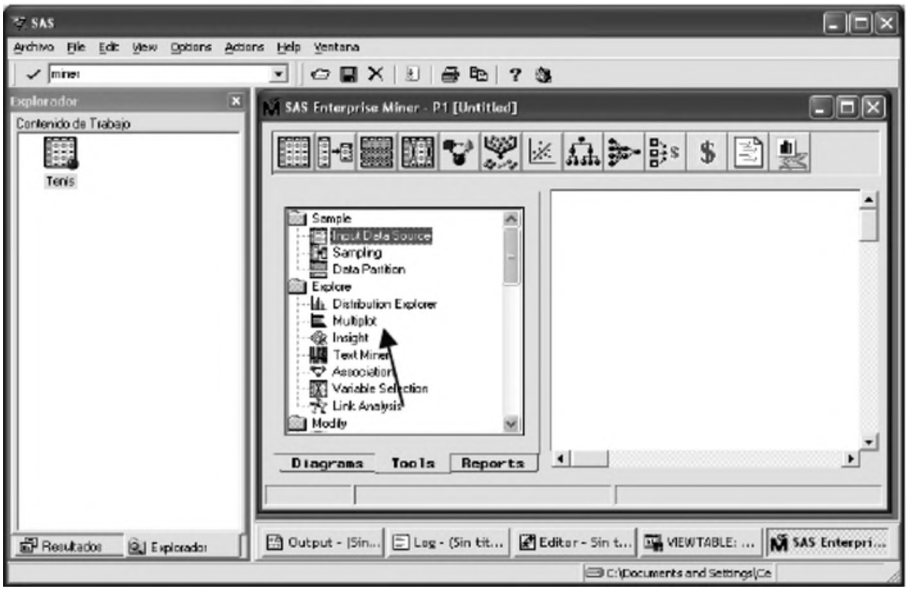


Figura 8-46

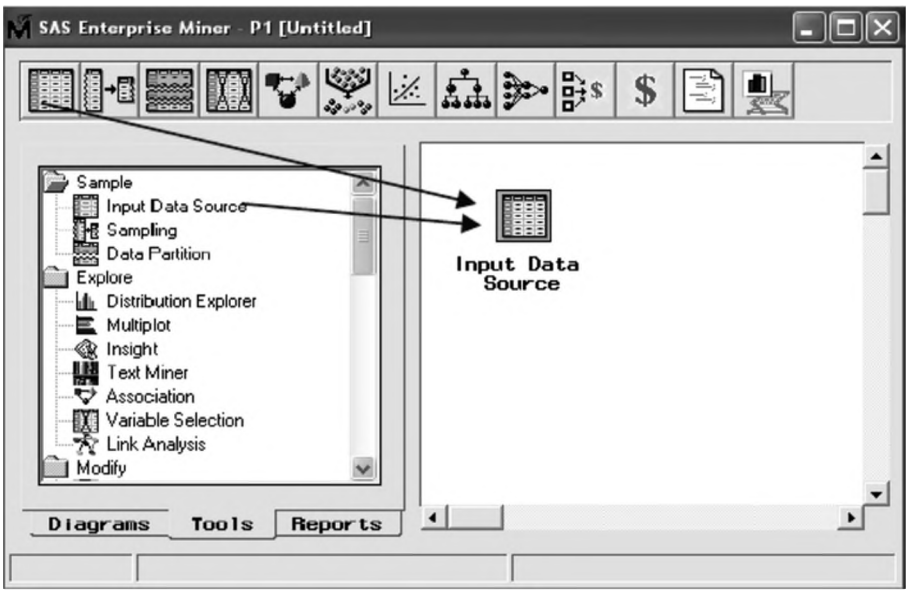


Figura 8-47

Al hacer doble clic sobre el nodo *Input Data Source* en la zona de trabajo, se obtiene la pantalla de entrada de datos. Se hace clic en la pestaña *Data* y en el campo *Source Data* debemos indicar en qué librería se encuentran los datos. Una vez que estos han sido seleccionados el nodo ofrece información acerca del número de filas y columnas que contiene el archivo (Figura 8-48). Además, el nodo realiza una valoración de las características de las variables. Si la muestra es muy grande, y dado que el objetivo perseguido es únicamente definir el tipo de variable, *Enterprise Miner* tomará una muestra que por defecto se fija en 2000 datos. Si la base de datos contiene menos que esta cantidad entonces se seleccionan todos los datos. De esta forma ya están enlazados los datos del fichero de entrada con *SAS Enterprise Miner*.

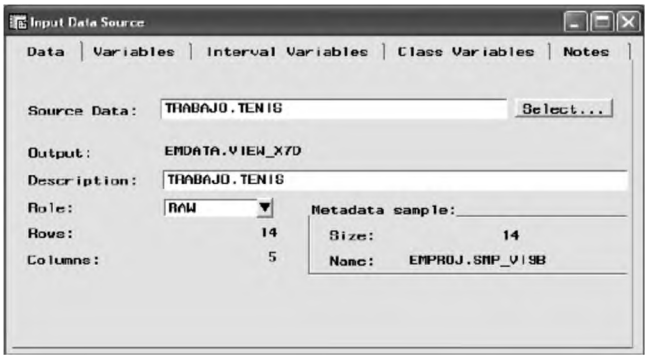


Figura 8-48

Nodo Input Data Source y los tipos de variables

La pestaña *Variables* de la pantalla *Input Data Source* (Figura 8-49) permite asignar tipos de variables siguiendo una serie de patrones que posteriormente pueden ser cambiados por el usuario. Nuestro problema es ver si hoy podemos jugar al tenis. Para poder abordar este problema hemos de definir las variables *Cielo, Temperatura, Humedad* y *Viento* como predictoras o independientes (de entrada o *input*), mientras que el campo *Jugar* es la clase a predecir (variable dependiente), o sea, el resultado (de destino o *target*). Por defecto *Enterprise Miner* define todas las variables como *input* y por tanto las variables *target* o variables respuesta deben ser asignadas por el usuario. Para ello se hace clic con el botón secundario del ratón sobre la variable a cambiar de tipo y se elige *Set Model Rol* en el menú emergente resultante (Figura 8-50). A continuación se elige el nuevo tipo (*Target*) en la lista de la Figura 8-51. La Figura 8-52 muestra las variables con su tipo definitivo para este problema. Por tanto el objetivo del nodo *Input Data Source* es tanto abrir los datos como definir el tipo de variable y su papel en el análisis de forma conveniente.

 ©Alfaomega-RC Libros

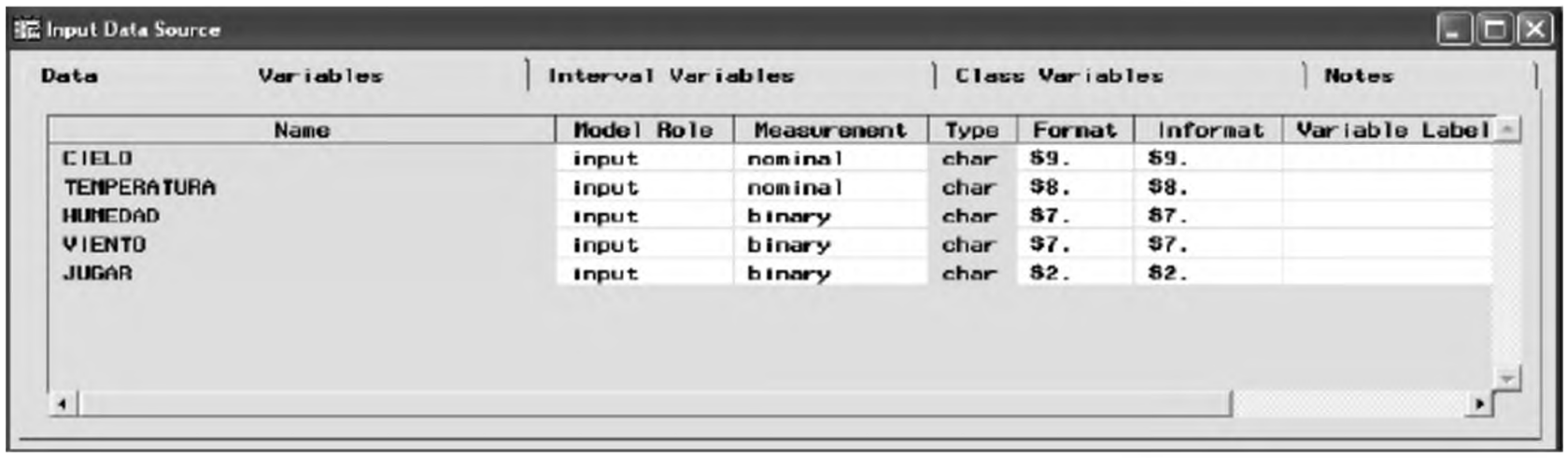


Figura 8-49

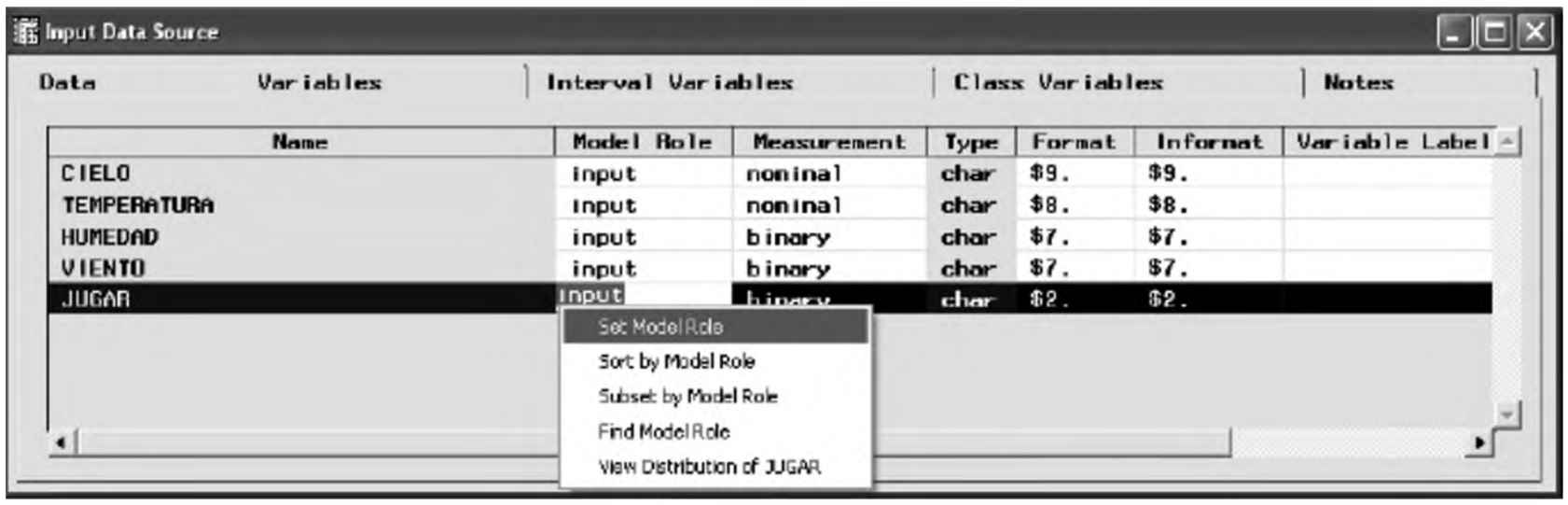


Figura 8-50

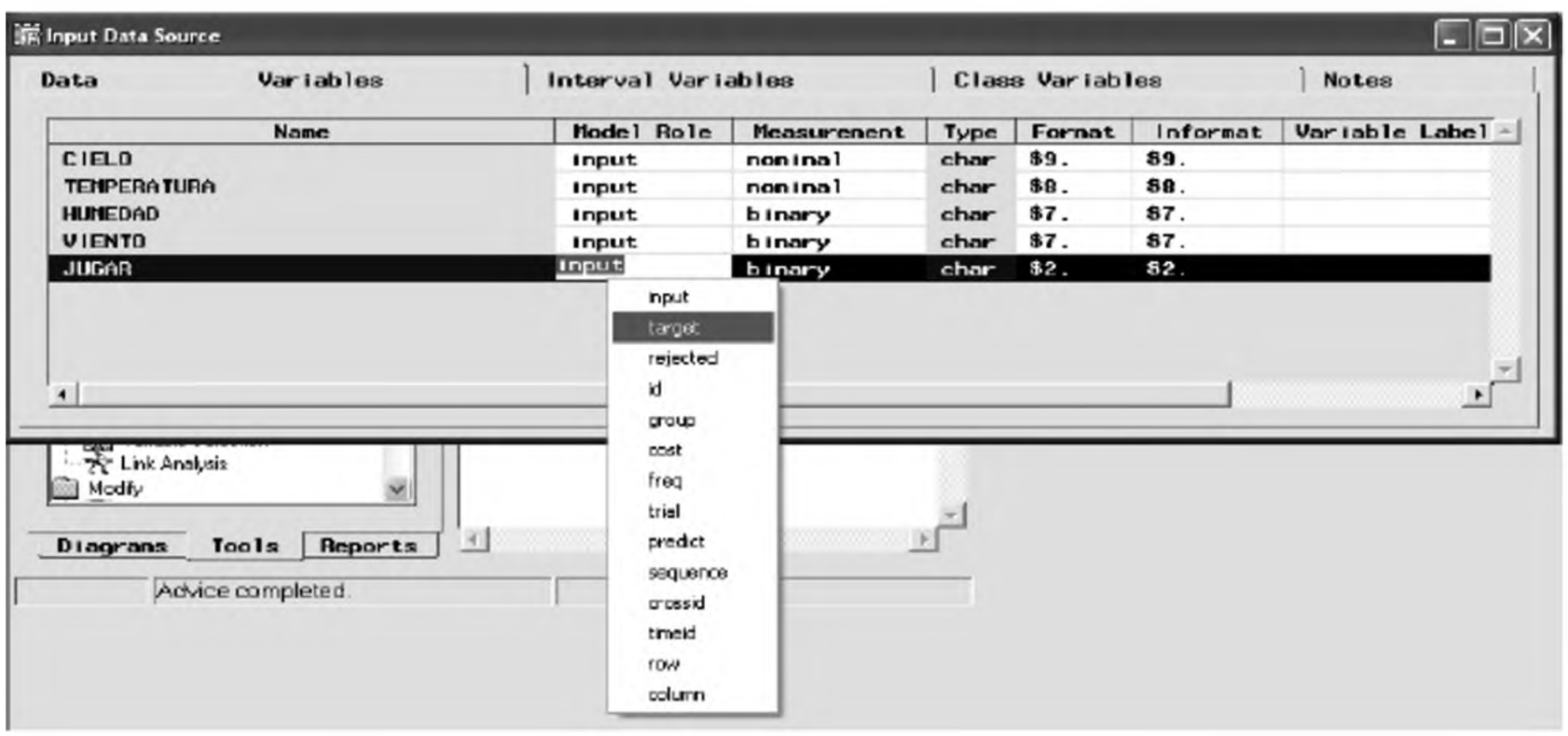


Figura 8-51

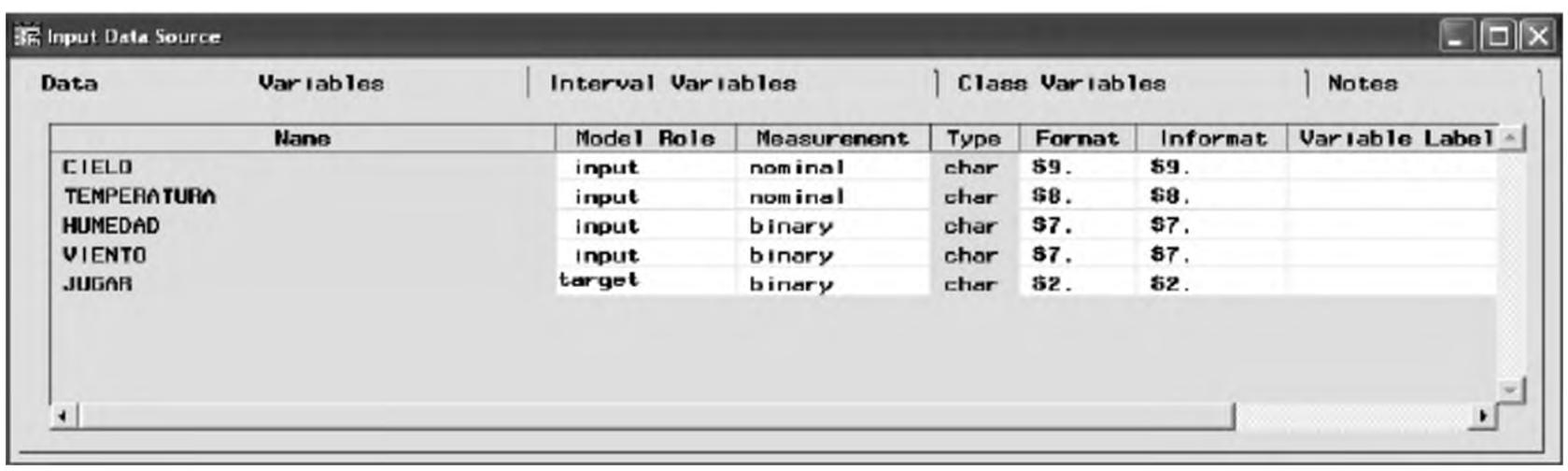


Figura 8-52

Enlace de nodos de un diagrama. El nodo Data Partition

Una vez que el nodo de datos recopila la información del problema, este debe ser conectado con otros nodos para implementar el proceso de *Data Mining*. La conexión de distintos nodos creará un diagrama lógico de flujo (*process flow*) que permitirá ir completando las etapas de muestreo, estadísticas descriptivas, codificación, modelado y valoración de resultados. La mayoría de los nodos solo pueden ser abiertos una vez que han sido conectados con nodos de datos.

En nuestro caso uniremos el nodo de datos (*Input Data Source*) con el de partición (*Data Partition*) con la finalidad de asignar de manera aleatoria un 80% de los datos a entrenamiento, un 10% a validación y un 10% a test. Para ello arrastramos el nodo *Data Partition* desde el listado de herramietas (*Tools*) o desde el icono a la zona de trabajo (Figura 8-53).

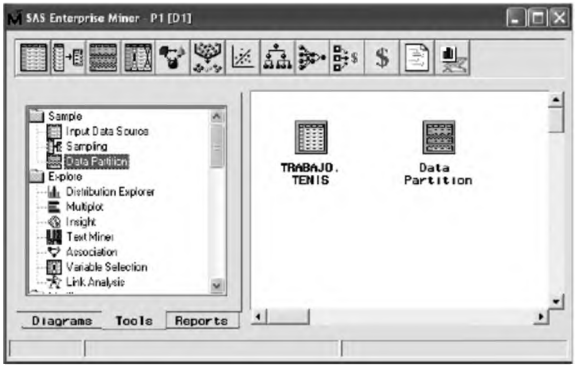


Figura 8-53

Para conectar dos nodos basta con hacer clic con el botón derecho sobre el nodo que queremos conectar y seleccionar *connect items* (Figura 8-54). Una vez elegida esta opción pulsaremos sobre el nodo origen y arrastraremos la flecha que se origina hasta el nodo destino (Figura 8-55).

Para borrar una conexión pulsaremos con el botón derecho del ratón sobre la flecha que conecta los dos nodos y seleccionaremos la opción *Delete* (Figura 8-56).

©Alfaomega-RC Libros

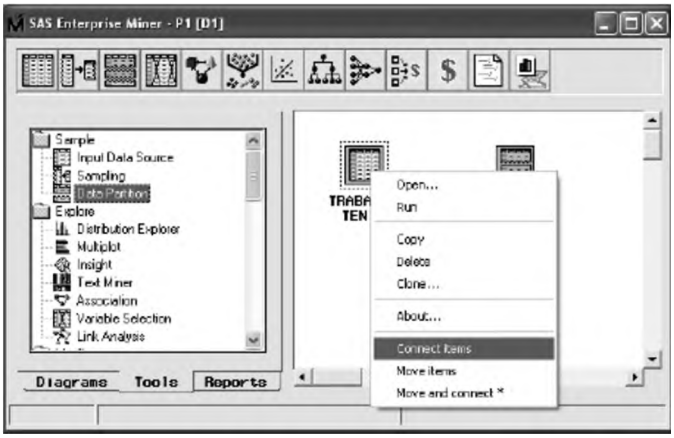


Figura 8-54

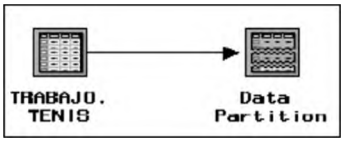


Figura 8-55

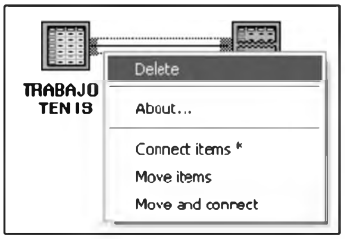


Figura 8-56

A medida que los nodos sean creados será necesario ubicarlos de forma adecuada sobre la zona de trabajo para lograr una mejor organización y comprensión del diagrama de flujo. Para mover un nodo pulsaremos sobre el botón derecho sobre el nodo que queremos desplazar y seleccionaremos la opción *Move items* de la Figura 8-54. Al elegir esta opción cuando situemos el cursor encima del nodo veremos cómo este toma la forma de una mano. Cuando esto ocurra podemos elegir el nodo con el botón izquierdo del ratón y sin soltar el nodo moverlo hasta la nueva posición. En general resulta conveniente seleccionar la opción *Move and connect*. Así, cuando el cursor esté sobre el borde del nodo, se transformará en una cruz indicando que podemos conectar dos nodos. Si situamos el cursor sobre el centro del ratón, podremos mover el nodo seleccionado.

Para asignar de manera aleatoria un 80% de los datos a entrenamiento, un 10% a validación y un 10% a test hacemos doble clic sobre el nodo *Data Partition* en la zona de trabajo y rellenamos la pestaña *Partition* como se indica en la Figura 8-57.

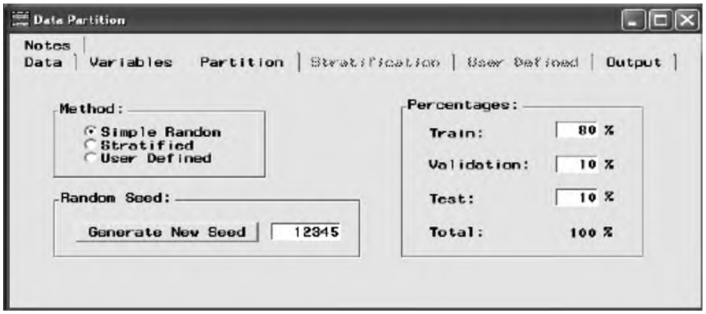


Figura 8-57

Nodo de modelado para crear el árbol

Ahora ya estamos en situación de intentar utilizar un modelo a partir de los datos, en nuestro caso un árbol de decisión. Para ello conectaremos el nodo de partición de datos con un nodo de árboles de decisión (*Tree*) que arrastramos a la zona de trabajo desde el listado de herramientas o desde el icono según se indica en la Figura 8-58.

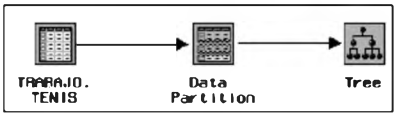


Figura 8-58

Una vez abierto el nodo *Tree* haciendo doble clic sobre él en el diagrama, en la pestaña *Advanced* especificaremos como medida para la valoración del modelo el porcentaje de casos correctamente clasificados. Además, definiremos como número de observaciones suficientes para buscar una variable de división 14 (Figura 8-59).

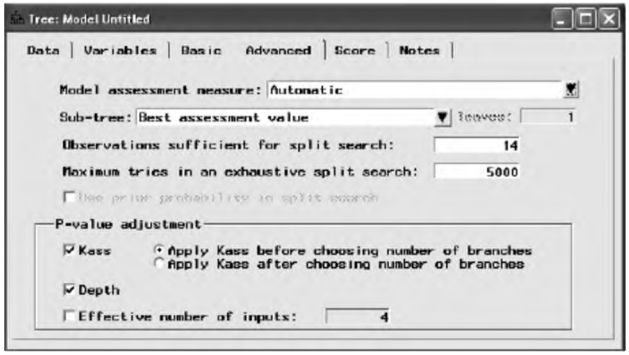


Figura 8-59

©Alfaomega-RC Libros

En la pestaña *Score* indicaremos que guarde la respuesta para los datos de entrenamiento, validación y test (Figura 8-60).

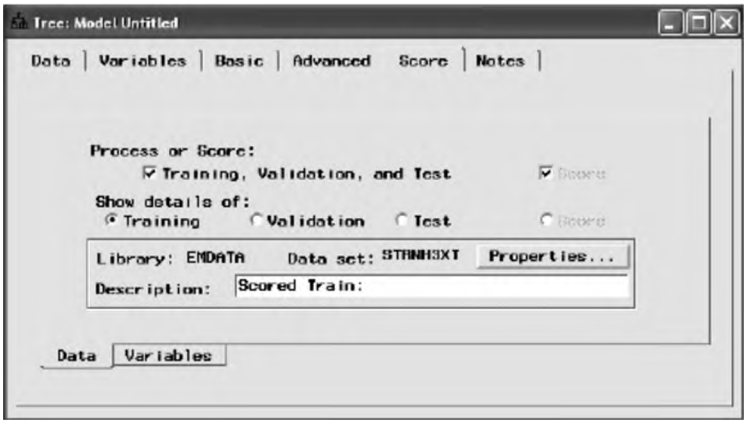


Figura 8-60

Una vez definidas estas opciones, y previa asignación de un nombre para el modelo (Figura 8-61), ejecutaremos el nodo mediante *Actions* → *Run* (Figura 8-62) o haciendo clic con el botón derecho del ratón sobre el nodo *Tree* y eligiendo *Run* (Figura 8-63). Después de una visión de ejecución con éxito (Figura 8-64), podemos ver los resultados del ajuste (Figura 8-65) que se interpretarán más adelante.

Figura 8-61

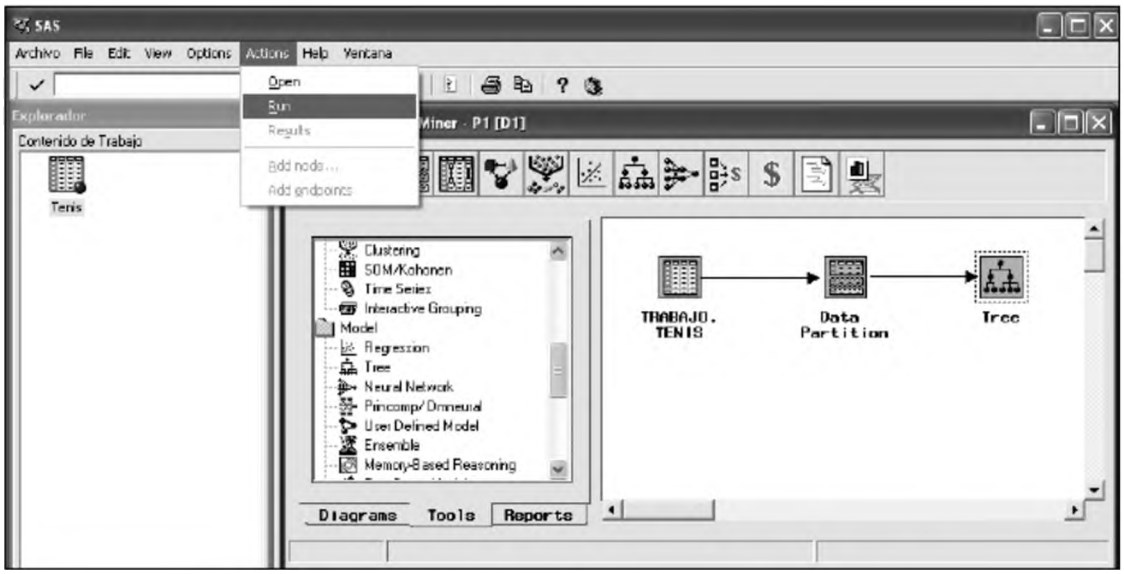


Figura 8-62

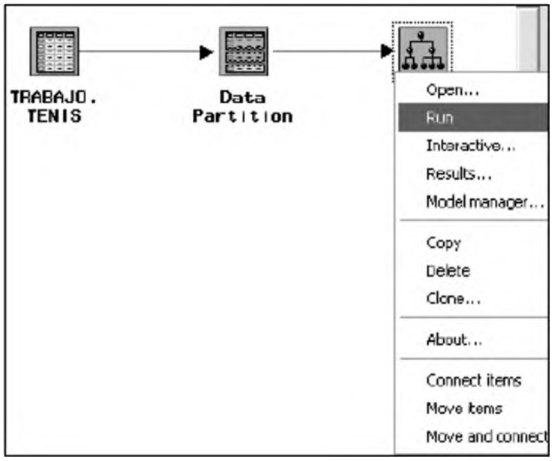


Figura 8-63

Figura 8-64

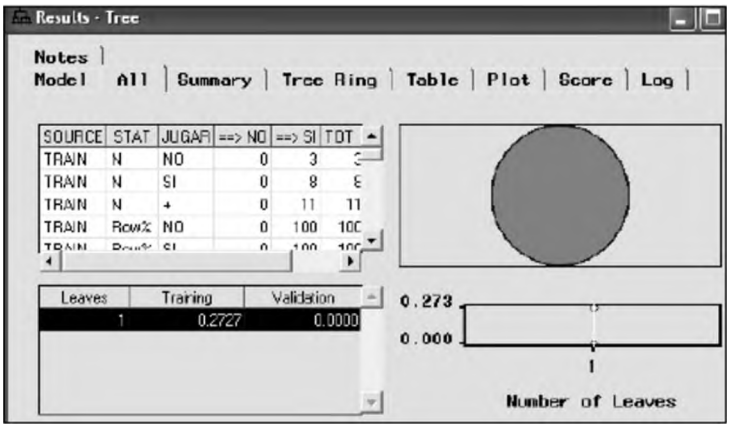


Figura 8-65

Podemos visualizar el árbol seleccionando *View* → *Tree* (Figura 8-66). Dado que el conjunto de datos es muy pequeño, se ve una sola rama (Figura 8-67).

Para solucionar el problema de una sola rama, eliminamos la fase intermedia de partición del conjunto de datos y construimos el árbol solamente con los nodos *Input Data Source* y *Tree*. Para ello abrimos un nuevo diagrama dentro de nuestro proyecto mediante *File* →*New* →*Diagram* (Figura 8-68). A continuación, insertamos en el diagrama los dos nodos y los unimos (Figura 8-69). Al ejecutar ahora el árbol obtenemos los resultados de la Figura 8-70.

©Alfaomega-RC Libros

La pestaña *All* muestra a la vez las cuatro pestañas siguientes. En primer lugar muestra las estadísticas de ajuste del modelo (pestaña *Summary*). En segundo lugar muestra la proporción de casos que se encuentra en cada uno de los nodos en cada uno de los niveles o anillos del árbol (pestaña *Ring*). En tercer lugar muestra la proporción de casos correctamente clasificados, tanto de entrenamiento como de validación, según el número de hojas de cada árbol (pestaña *Table*). Por último se realiza una representación gráfica del resultado anterior (pestaña *plot*).

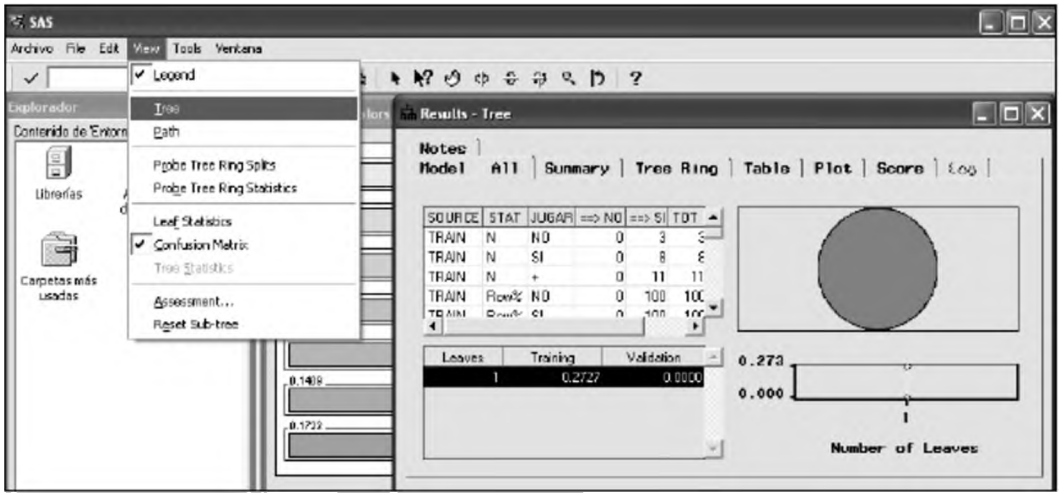


Figura 8-66

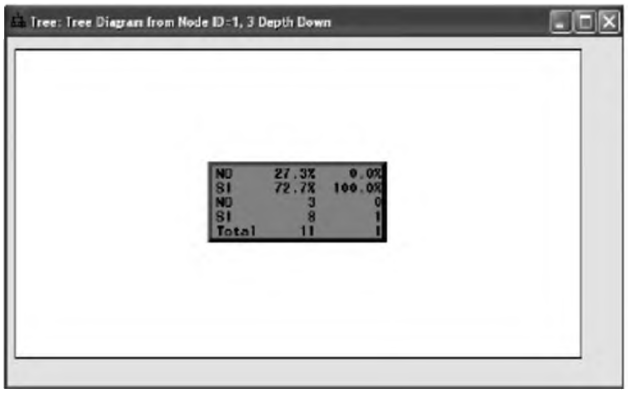


Figura 8-67

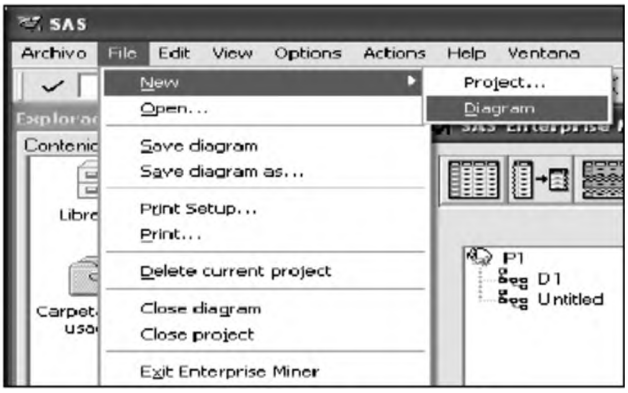


Figura 8-68

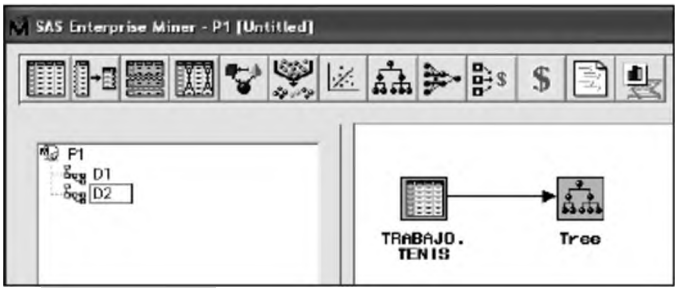


Figura 8-69

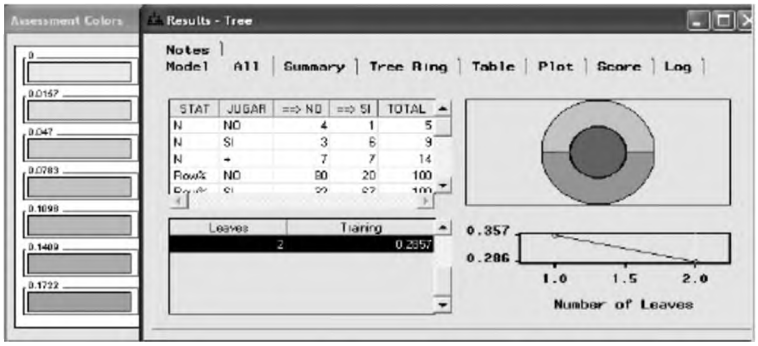


Figura 8-70

Podemos visualizar el árbol seleccionando *View* → *Tree* (Figura 8-71).

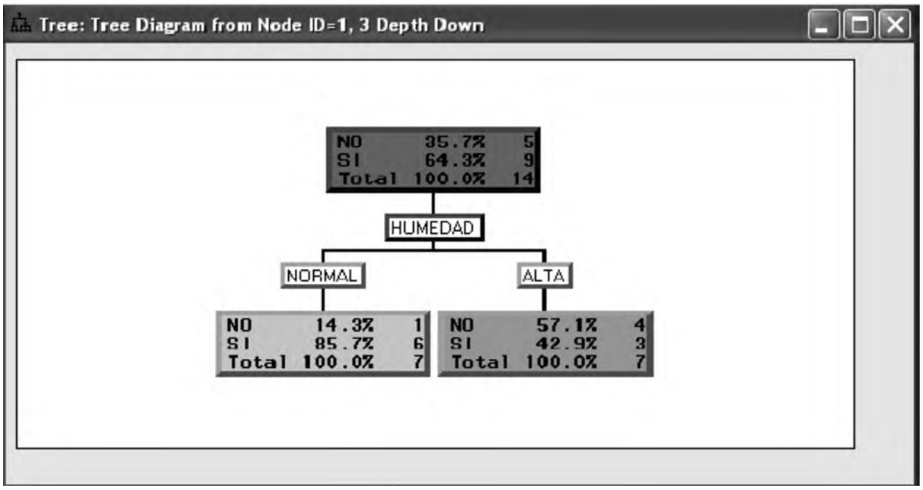


Figura 8-71

PROFUNDIZANDO EN EL NODO TREE

El nodo *Tree* de Enterprise Miner (Figura 8-72) permite implementar distintos algoritmos para la generación de reglas lógicas de decisión elaboradas a partir de la información contenida en los datos. Su uso requiere una base de datos en la cual debemos especificar una única respuesta, que puede ser nominal, binaria, ordinal o continua, y al menos un *input* o variables explicativas. El nodo *Tree* se encuentra situado en SAS Enterprise Miner en la categoría *Model* (Figura 8-73).

Figura 8-72

©Alfaomega-RC Libros

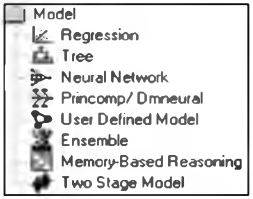


Figura 8-73

Como ejemplo, vamos a utilizar los datos del archivo *Créditos* para establecer las reglas lógicas que van a permitir ayudar a tomar la decisión sobre cuándo decidir asignar o no un crédito. Una vez importado el fichero, asignaremos a la variable *Cliente* el papel ID en el modelo mientras que la variable *CREDIT_V* será la variable objetivo (*Target*) y el resto de variables serán *inputs*.

Una vez que disponemos ya de nuestros datos de trabajo en formato SAS (archivo *creditos.sas7bdat*) en una determinada librería (librería *Trabajo* que representa el subdirectorio *c:\libros\miningt*), abrimos el proyecto P1 (*File* → *Open*) y mediante *File* → *New* → *Diagram* creamos el diagrama D13. A continuación, situamos en el diagrama el nodo *Input Data Source* al que se le asigna el conjunto de datos *creditos.sas7bdat* de la librería TRABAJO. A continuación uniremos el nodo de datos con el de partición (*Data Partition*) y asignaremos de manera aleatoria un 80% de los datos a entrenamiento, un 10% a validación y un 10% a test. Una vez realizada esta tarea conectaremos el nodo de partición de datos con un nodo de árboles de decisión (*Tree*) según se indica en la Figura 8-74.

Para asignar a la variable *Cliente* el tipo *id*, a la variable *CREDIT_V* el tipo *Target* y al resto de variables el tipo *input*, hacemos doble clic sobre el nodo *Input Data Source* y elegimos la pestaña *Variables*. A continuación se hace clic con el botón derecho del ratón sobre cada variable a situar de tipo *input*, se elige *Set Model Role* y a continuación *Input*. Se repite el proceso con las variables de tipo *id* y *target* (Figura 8-75). Para asignar de manera aleatoria un 80% de los datos a entrenamiento, un 10% a validación y un 10% a test hacemos doble clic sobre el nodo *Data Partition* y rellenamos la pestaña *Variables* como se indica en la Figura 8-76.

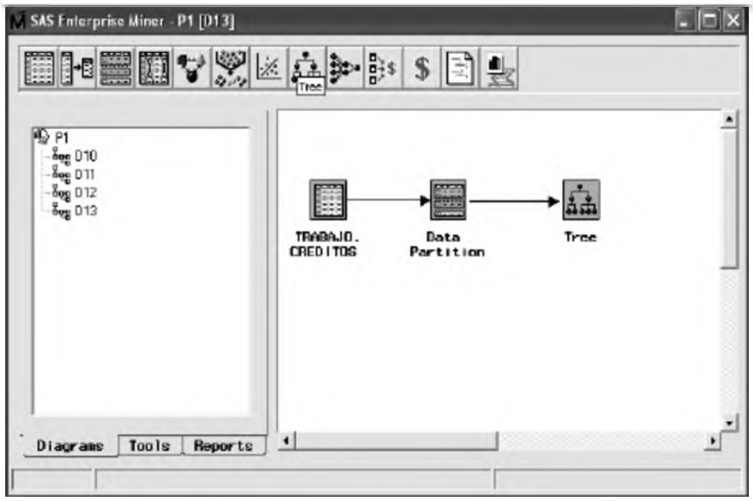


Figura 8-74

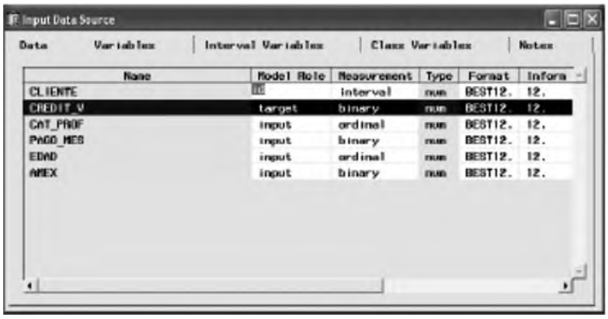


Figura 8-75

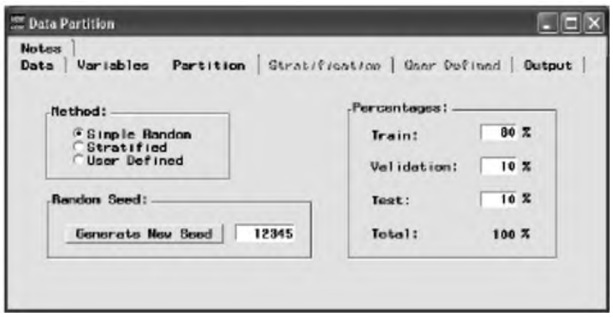


Figura 8-76

Para fijar las especificaciones del árbol abrimos el nodo *Tree* haciendo doble clic sobre él en el diagrama. Se obtiene la pantalla de entrada del nodo en cuya pestaña *Variables* se presentan las variables que intervienen en el árbol y sus propiedades (Figura 8-77).

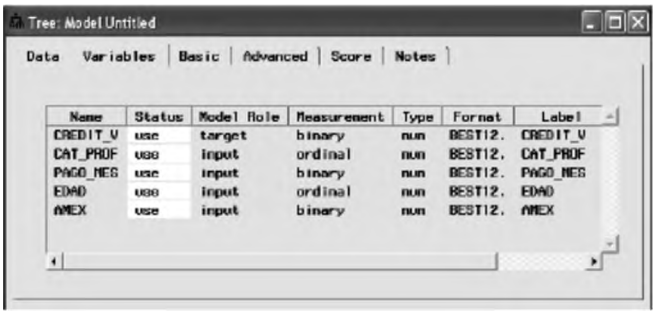


Figura 8-77

En la pestaña *Advanced* especificaremos como medida para la valoración del modelo el porcentaje de casos correctamente clasificados. Además, definiremos como número de observaciones suficientes para buscar una variable de división 25 (Figura 8-78).

©Alfaomega-RC Libros

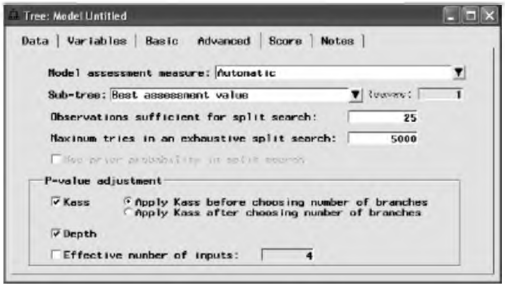


Figura 8-78

En la ventana de opciones avanzadas obtenida con la pestaña *Advanced* (Figura 8-78). Se pueden especificar las siguientes variables:

Model Assessment Measure: selecciona el mejor árbol a partir de los resultados en la muestra de validación. La lista de medidas de valoración depende de cómo el *output* esté medido y de si existe o no una matriz de costes o beneficios para el objetivo.

Para variables continuas se puede elegir entre varias medidas dependiendo de si existe o no una matriz de beneficios. Si no existe matriz de beneficios tendremos:

- *Average Square Error:* medida por defecto.
- *Average in the top 10, 25 or 50%.*

Si existe matriz de beneficios o pérdidas tendremos:

- *Average Square Error.*
- *Average profit/loss.*
- *Average profit/loss in the top 10, 25 or 50%.*

Para objetivos categóricos podemos elegir entre las siguientes medidas:

Si no existe matriz de pérdidas o beneficios definida:

- *Proportion correctly classified:* proporción de casos clasificados correctamente.
- *Ordinal proportion correctly classified.*
- *Proportion of event in top 10, 25 or 50%.*
- *Total leaf Impurity (Fini index).*

Si existe una matriz de pérdidas o beneficios entonces las medidas de valoración serán:

- *Proportion correctly classified.*

- *Average profit/loss.*
- *Average profit/loss in top 10, 25 or 50%.*
- *Total leaf impurity (Gini index).*

Specify the subtree method: esta opción especifica cómo seleccionar un subárbol dentro del árbol principal. Si un árbol tiende a tener demasiadas ramas y demasiada profundidad tenderá también a ajustar el ruido y generará de forma pobre nuevos datos que le sean presentados. Cada nodo puede ser evaluado de las siguientes formas:

- ***Best assessment value****:* el subárbol más pequeño con la mejor valoración es el elegido por defecto.
- ***The most leaves****:* selecciona el árbol entero. Esta opción es apropiada cuando el árbol se construye de forma interactiva o cuando se definen otras opciones para parar las divisiones del árbol.
- ***At most indicated number of leaves****:* selecciona el subárbol que tenga como mucho *n* hojas. Este número debe ser definido en el cuadrado correspondiente que se activará al elegir esta opción (Figura 8-79).

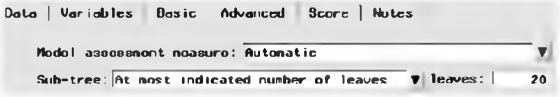


Figura 8-79

Split search criteria*:* esta opción establece un límite superior en el número de observaciones de la muestra para realizar una partición. El algoritmo para la creación de un árbol de decisión busca reglas de partición en ramas que maximicen el criterio asignado. Encontrar la partición óptima supone a menudo evaluar cada posible división de cada variable y a veces el número de posibles divisiones puede llegar a ser excesivamente grande.

Maximum tries in an exhaustive split search*:* si el número de posibles divisiones es elevado, esta opción utiliza un algoritmo de búsqueda (*stepwise, hill-climbing*) con un número de intentos predeterminado.

P-value adjustment*:* esta opción se activa cuando se elige como criterio de valoración los valores de los estadísticos chi-cuadrado o F.

Kass*:* la búsqueda de la partición óptima requiere el cálculo de distintas tablas de contingencia. Si usamos la tabla original sin cambios en las categorías, el test χ^2 puede ser usado. Este test asume que solo existe una población de la que extraemos una única muestra y calculamos un único test. Sin embargo, realizar el test de forma

 ©Alfaomega-RC Libros

repetida viola este supuesto. Ello aumenta la posibilidad de encontrar alguna relación simplemente por el hecho de incrementar el número de veces en la búsqueda lo que puede llevar a encontrar relaciones espurias o a magnificar las relaciones encontradas. Es por ello que el *p*-valor es corregido de la forma descrita en la descripción realizada de los árboles de decisión.

Para implementar el algoritmo, el primer paso es crear regresores ordinales en lugar de continuos dividiendo la función de distribución de las variables explicativas continuas en un número de categorías que contenga aproximadamente el mismo número de observaciones. Para las variables categóricas, las categorías permanecen ya definidas. Esta sería la opción ***Apply Kass before choosing number of branches***.

Sin embargo, esta opción puede rebajar más la significación de una partición que un método alternativo denominado ajuste de Gabriel que se aplica si se selecciona ***Apply Kass after choosing number of branches***.

Depth: esta opción realiza un ajuste de Bonferroni teniendo en cuenta el número de hojas para corregir el número de rechazos falsos que tenderá a crecer con el número de hojas.

Effective number of inputs: ajusta los p-valores a partir del número efectivo de *inputs*. Cuantos más inputs más probable será que un *input* espurio gane al *input* o *inputs* verdaderamente predictivos. Cuantos más *inputs* incorrelacionados haya más alto será el riesgo. El ajuste por *inputs* multiplica el *p*-valor por el número declarado de *inputs* que son aquellos que tienen el estatus de uso en la ventana de variables.

Si ahora en la pantalla *Tree node* de especificaciones del árbol elegimos la pestaña *Score*, podemos indicar que se guarde la respuesta para los datos de entrenamiento, validación y test (Figura 8-80).

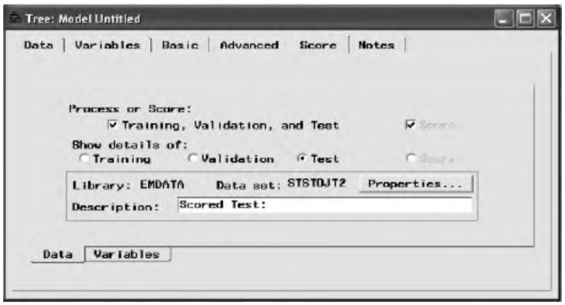


Figura 8-80

Se observa que la pestaña *Score* dispone de dos subventanas:

***Data**:* sirve para seleccionar los datos para valorar (*score*) cuando el nodo de árboles de decisión sea ejecutado (Figura 8-80). Por defecto ni los datos de entrenamiento, ni de validación ni de test son usados para valorar. Para que sean usados con este fin es necesario activar el cuadrado de *Training, Validation and Test.* En esta ventana también se pueden ver los detalles acerca de los datos de entrenamiento, validación y test seleccionando la opción *Properties*.

Variables: en esta ventana (Figura 8-81) se seleccionan aquellas variables que deseamos estén disponibles para futuros análisis en otros nodos.

Figura 8-81

Se pueden seleccionar las siguientes variables:

***Input variable selection**:* preselecciona las variables predictoras importantes reduciendo de esta forma la dimensionalidad de los datos. Cuando se ejecuta el nodo de árboles de decisión asignará el estatus de *input* a aquellas variables con un nivel de confianza mayor del 95% asignando el estatus de rechazada al resto. Esto permitirá incluir estas variables en un nodo posterior, como por ejemplo de redes neuronales.

***New variables related to score**:*

- ***Dummy variables**:* crea una variable *dummy* para cada hoja del árbol. Si cada observación está asignada a una hoja específica entonces tendrá asociado un valor de 1 para esa hoja y de 0 para el resto.
- ***Leaf identification variable**:* esta variable contiene una identificación numérica de la hoja en la cual la observación es asignada. Puede ser usada para procesar grupos en un nodo de procesado de grupos o realizar regresiones para cada uno de los grupos encontrados.
- ***Prediction variables**:* variables de predicción que pueden ser usadas en análisis sucesivos.

©Alfaomega-RC Libros

- Una vez establecidas todas las opciones de entrenamiento el nodo puede ser ejecutado de cuatro formas:

  Si el nodo está cerrado:

- Seleccionar el nodo con el botón derecho y seleccionar *Run*.
- Seleccionar *Actions* → *Run*.

  Si el nodo está abierto:

- Seleccionar *Tools* → *Train Model*.
- Seleccionar el icono para ejecutar el árbol .

Como veremos posteriormente el entrenamiento del árbol también puede ser ejecutado de forma interactiva.

Una vez definidas estas opciones, y previa asignación de un nombre para el modelo (Figura 8-82), ejecutaremos el nodo haciendo clic sobre él en el diagrama con el botón derecho del ratón y eligiendo la opción *Run* del menú emergente resultante (Figura 8-83). Finalizada la ejecución el sistema pregunta si queremos ver los resultados (Figura 8-84). Al aceptar se abre el *visor de resultados* cuya pestaña *All* (Figura 8-85) muestra a la vez cuatro pestañas.

En la parte superior izquierda, la pestaña *Summary* muestra las estadísticas de ajuste del modelo tanto para los datos de entrenamiento como de validación. En la parte superior derecha, la pestaña. *Ring* muestra la proporción de casos que se encuentra en cada uno de los nodos en cada uno de los niveles o anillos del árbol. Seleccionando en la barra de herramientas el icono *View Information About Point* podemos ver la regla que define el nodo (Figura 8-86). En la parte inferior izquierda, la pestaña *Table* muestra la proporción de casos correctamente clasificados, tanto de entrenamiento como de validación, según el número de hojas de cada árbol. En la parte inferior derecha, la pestaña *Plot* realiza una representación gráfica del resultado anterior. En un árbol de decisión interesa elegir el punto que presenta un mayor número de casos bien clasificados en la muestra de validación. En este caso nos quedaríamos con el árbol de 5 hojas porque la línea blanca vertical está en el 5. Puede cambiarse de punto haciendo clic en el diamante de la figura.

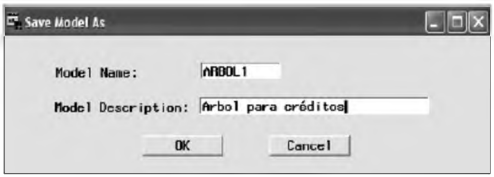


Figura 8-82

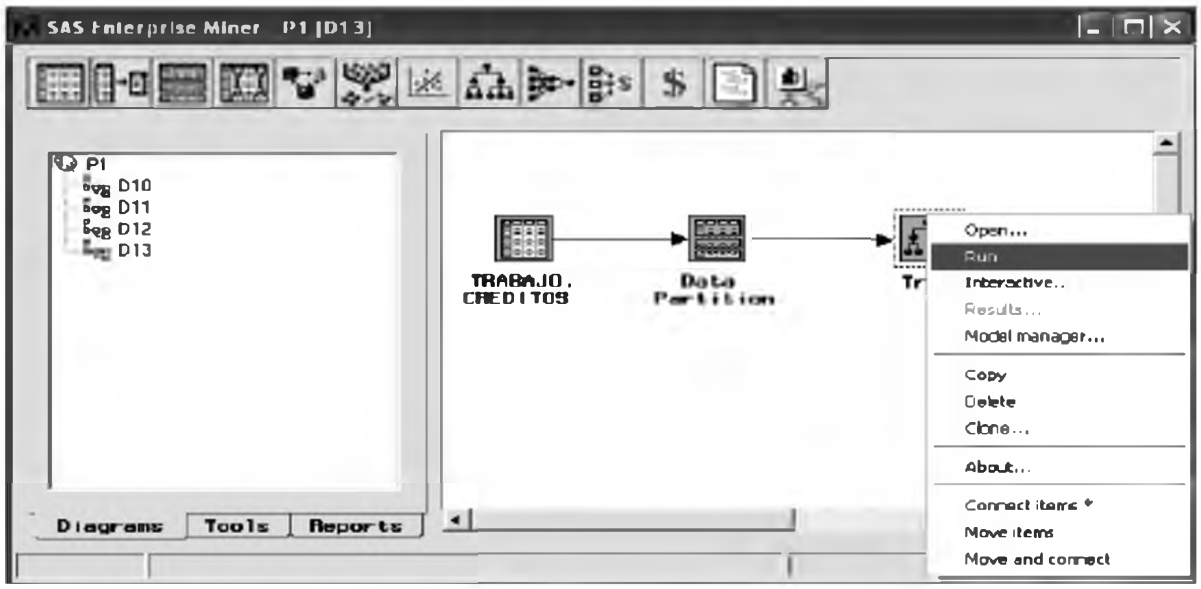


Figura 8-83

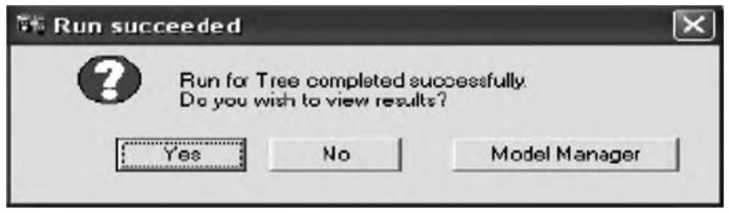


Figura 8-84

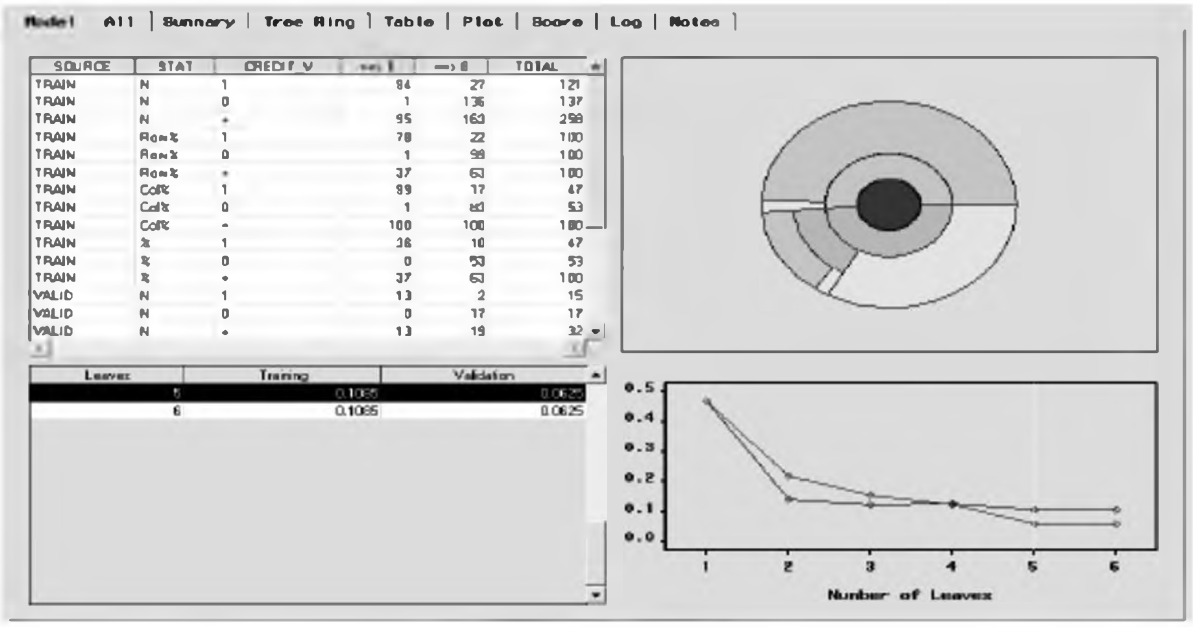


Figura 8-85

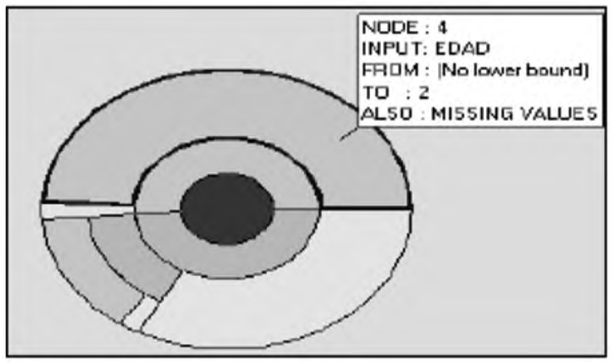


Figura 8-86

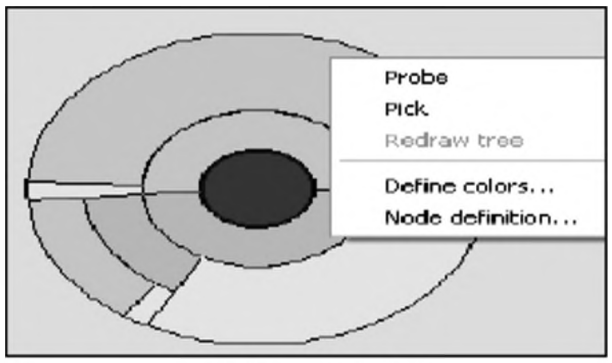


Figura 8-87

Dentro de la ventana del gráfico del anillo del árbol Tree Ring, si pulsamos con el botón derecho del ratón sobre el gráfico (Figura 8-87), podemos seleccionar las siguientes opciones:

©Alfaomega-RC Libros

- ***Probe:*** muestra cierta información resumen en un cuadro de texto al seleccionar un nodo o simplemente moviendo el cursor por encima del gráfico (Figura 8-86).
- ***Pick:*** selecciona y destaca un nodo sobre los demás (Figura 8-86). Una vez seleccionado haremos clic en *Node Definition* para evaluar la regla lógica que define sus propiedades.
- ***Redraw tree:*** vuelve a dibujar el diagrama del árbol de decisión clásico a partir del nodo seleccionado. Esta opción es especialmente útil para visualizar detalles del árbol ya que normalmente el número de ramas, nodos y la profundidad del árbol harán que sea imposible su visualización completa en una sola ventana. Para imprimir el árbol entero en ocasiones será necesario guardar la imagen y posteriormente modificarla en otro programa que permita su manipulación.
- ***Define colors:*** elige los colores deseados según preferencias para resaltar determinados resultados (Figura 8-88).

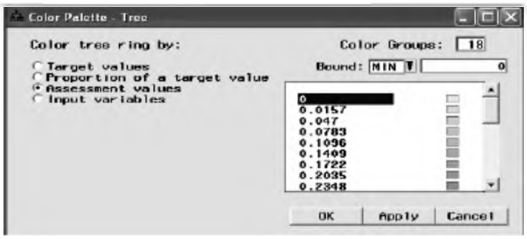


Figura 8-88

- ***Node definition:*** muestra la definición del nodo (regla lógica en lengua inglesa) para el nodo seleccionado.

Para *visualizar el árbol* seleccionamos *View* →*Tree* (Figura 8-89). La columna del medio muestra los porcentajes y valores numéricos de los datos de entrenamiento mientras que a la derecha se muestran los datos de validación (Figura 8-90).

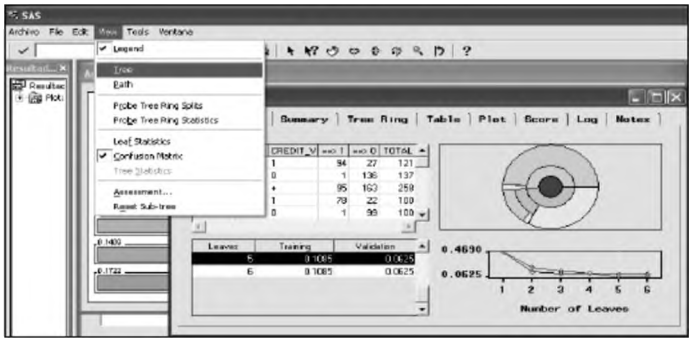


Figura 8-89

Un diagrama de árbol de decisión contiene las siguientes características:

- *Nodo Raíz:* es el nodo que contiene todas las observaciones justo antes de la primera división. En este caso se muestra que de los 258 clientes, 121 devolvieron el crédito (46,9%) mientras que 137 resultaron impagados (53,1%).
- *Ramas:* una vez seleccionada la variable que mejor discrimina, los datos son partidos en dos o más ramas de acuerdo con los valores de la variable. En el caso anterior la variable que más discrimina es si el cliente percibe su remuneración de forma mensual (2) o semanal (1).
- *Nodos:* contienen los datos divididos procedentes de otras ramas y nodos.
- *Nodos hoja:* son los nodos terminales y contienen la clasificación general del árbol una vez cumplidas todas las reglas lógicas definidas por los nodos anteriores.

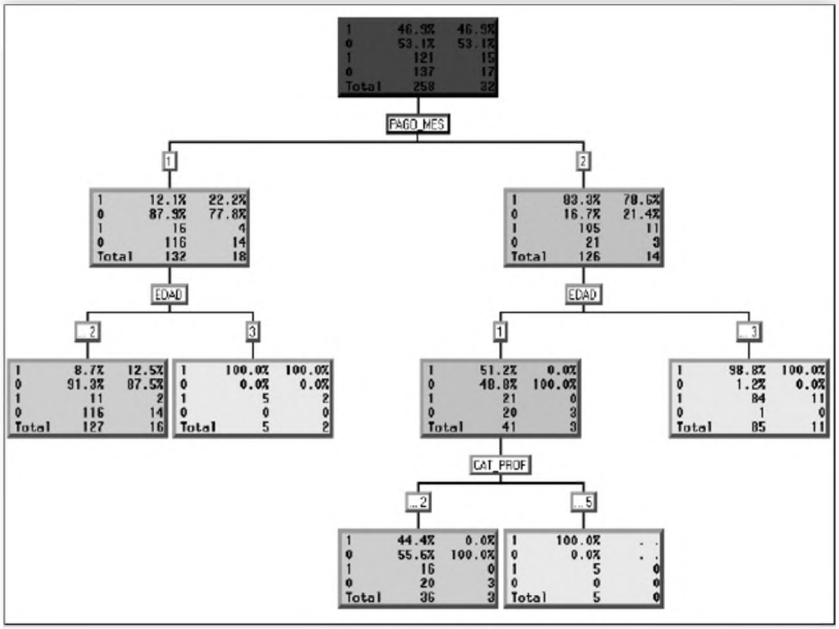


Figura 8-90

©Alfaomega-RC Libros

CAPÍTULO 9

REGLAS DE ASOCIACIÓN CON HERRAMIENTAS DE IBM Y SAS

REGLAS DE ASOCIACIÓN

Las reglas de asociación relacionan una determinada conclusión (la compra de un producto dado) con un conjunto de condiciones (la compra de otros productos). Por ejemplo, la regla

cerveza <= lata_veg & congelados (173, 17,0%, 0,84)

indica que, a menudo, se da el caso de cerveza cuando lata_veg y congelados ocurren al mismo tiempo. La regla es fiable en un 84% y se aplica al 17% de los datos (o 173 registros).

Los algoritmos de reglas de asociación buscan automáticamente las asociaciones que se podrían encontrar manualmente usando técnicas de visualización como en el nodo *Malla* (Figura 9-1).

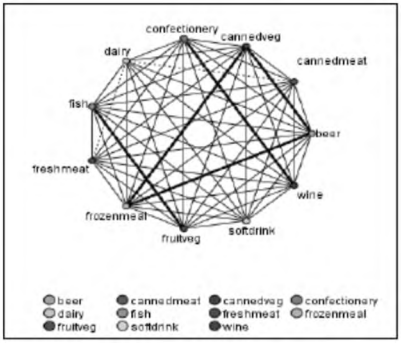


Figura 9-1

La ventaja de los algoritmos de reglas de asociación sobre los algoritmos más estándar de árboles de decisión (C5.0 y Árbol C&R) es que las asociaciones pueden existir entre cualquiera de los atributos. Un algoritmo de árbol de decisión generará reglas con una única conclusión, mientras que los algoritmos de asociación tratan de buscar muchas reglas, cada una de las cuales puede tener una conclusión diferente.

La desventaja de los algoritmos de asociación es que tratan de encontrar patrones en un espacio de búsqueda potencialmente muy amplio y, por tanto, pueden necesitar mucho más tiempo de ejecución que un algoritmo de árbol de decisión. Los algoritmos usan un método de generación y comprobación para buscar reglas: se generan inicialmente reglas sencillas que se validan basándose en el conjunto de datos. Las buenas reglas se almacenan y todas las reglas, sujetas a varias restricciones, se especializan posteriormente. La especialización es el proceso de añadir condiciones a una regla. Estas nuevas reglas se validan basándose en los datos y el proceso almacena de forma iterativa las mejores reglas o las más interesantes que se encuentren. El usuario proporciona generalmente alguna limitación al número posible de antecedentes que permitir en una regla, y se usan diversas técnicas basadas en la teoría de la información o esquemas de indización eficientes para reducir el potencialmente amplio espacio de la búsqueda.

Al final del procesamiento se presenta una tabla con las mejores reglas. A diferencia de un árbol de decisión, este conjunto de reglas de asociación no se puede usar directamente para realizar pronósticos de mismo modo que puede hacerlo un modelo estándar (como un árbol de decisión o una red neuronal). Esto se debe a las diversas conclusiones diferentes posibles de las reglas. Otro nivel de transformación es preciso para transformar las reglas de asociación en un conjunto de reglas de clasificación. Por tanto, las reglas de asociación producidas por algoritmos de asociación se conocen como modelos sin refinar. Aunque el usuario puede examinar estos modelos sin definir, estos no se pueden usar explícitamente como modelos de

 ©Alfaomega-RC Libros

clasificación a menos que el usuario indique al sistema que genere un modelo de clasificación a partir del modelo sin definir. Este se lleva a cabo desde el explorador a través de una opción del menú *Generar*.

En IBM SPSS Modeler se admiten tres algoritmos de reglas de asociación *A priori, Carma* y *Secuencia* situados en la categoría *Asociación* de la ficha *Modelado* (Figura 9-2).

Figura 9-2

REGLAS DE ASOCIACIÓN EN IBM SPSS. EL NODO A PRIORI

El nodo A priori extrae un conjunto de reglas de los datos y destaca aquellas reglas con un mayor contenido de información. A priori ofrece cinco métodos diferentes para la selección de reglas y utiliza un sofisticado esquema de indización para procesar eficientemente grandes conjuntos de datos. En los problemas de mucho volumen, A priori se entrena más rápidamente, no tiene un límite arbitrario para el número de reglas que puede retener y puede gestionar reglas que tengan hasta 32 precondiciones. A priori requiere que todos los campos de entrada y salida sean categóricos, pero ofrece un mejor rendimiento ya que está optimizado para este tipo de datos.

El nodo A priori encuentra reglas de asociación en los datos. Las reglas de asociación son instrucciones del tipo

si antecedente(s) entonces consecuente(s)

Por ejemplo, "si un cliente compra una cuchilla y una loción para después del afeitado, hay un 80% de posibilidades de que el cliente compre también crema de afeitado". A priori extrae un conjunto de reglas de los datos y destaca aquellas reglas con un mayor contenido de información. A priori ofrece cinco métodos diferentes para la selección de reglas y utiliza un sofisticado esquema de indización para procesar conjuntos de datos de gran tamaño de forma eficaz.

Para crear un conjunto de reglas de A priori, se precisan uno o varios campos de *Entrada* y uno o varios campos de *Objetivo*. Los campos de entrada y de salida (con dirección *Entrada, Objetivo* o *Ambos*) deben ser simbólicos. Los campos con el papel *Ninguno* se omiten. Los tipos de campo deben estar completamente instanciados antes de ejecutar el nodo. Los datos pueden estar en formato tabular o transaccional.

En los problemas de grandes dimensiones, *A priori* se entrena más rápidamente. Tampoco tiene un límite arbitrario para el número de reglas que puede retenerse y puede gestionar reglas que tengan hasta 32 precondiciones. A priori ofrece cinco métodos de entrenamiento distintos, lo que permite una mayor flexibilidad para asociar el método de minería de datos con el problema en cuestión.

Ilustramos el algoritmo con un ejemplo que busca las asociaciones entre los productos alimenticios de la dieta de jugadores de baloncesto. Las variables se recopilan en el archivo ASCII *baskets1n*.

Comenzamos utilizando como origen de datos el conjunto de datos ASCII en formato variable *baskets1n.txt* y lo unimos con un nodo *Tipo* para situar como variables de medida *Marca* y papel *Ambas* todas las involucradas en el análisis de asociaciones (variables de tipo alimenticio) y como no utilizables el resto de las variables del archivo (Figura 9-3).

Para realizar el análisis de asociación añadimos el nodo *A priori* de la categoría *Asociación* de la ficha *Modelado* al diagrama. A continuación hacemos clic con el botón derecho del ratón sobre el nodo *Asociación* y elegimos *Edición* en el menú emergente resultante. Se obtiene la pantalla de la Figura 9-4 en cuya solapa *Campos* elegimos *Utilizar la configuración del nodo tipo*. En la solapa *Modelo* elegimos *Automático, Utilizar los datos en particiones, Solo valores verdaderos para las marcas y Optimizar velocidad* (Figura 9-5). En al solapa *Experto* (Figura 9-6) elegimos *Experto* y *Medida de evaluación* por *Confianza de la regla*. Al hacer clic en *Aplicar* y *Ejecutar* se obtiene el nodo *A priori* sobre la ruta (Figura 9-6).

©Alfaomega-RC Libros

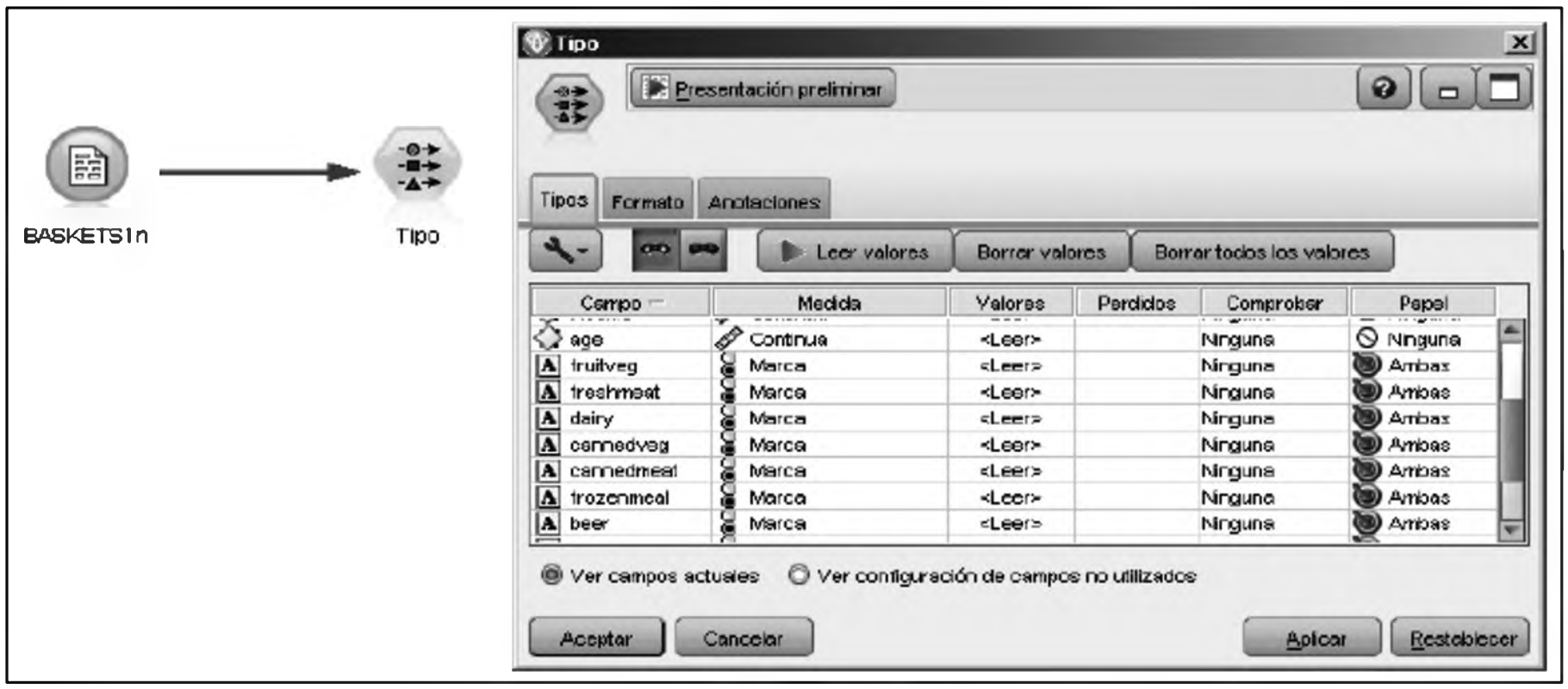


Figura 9-3

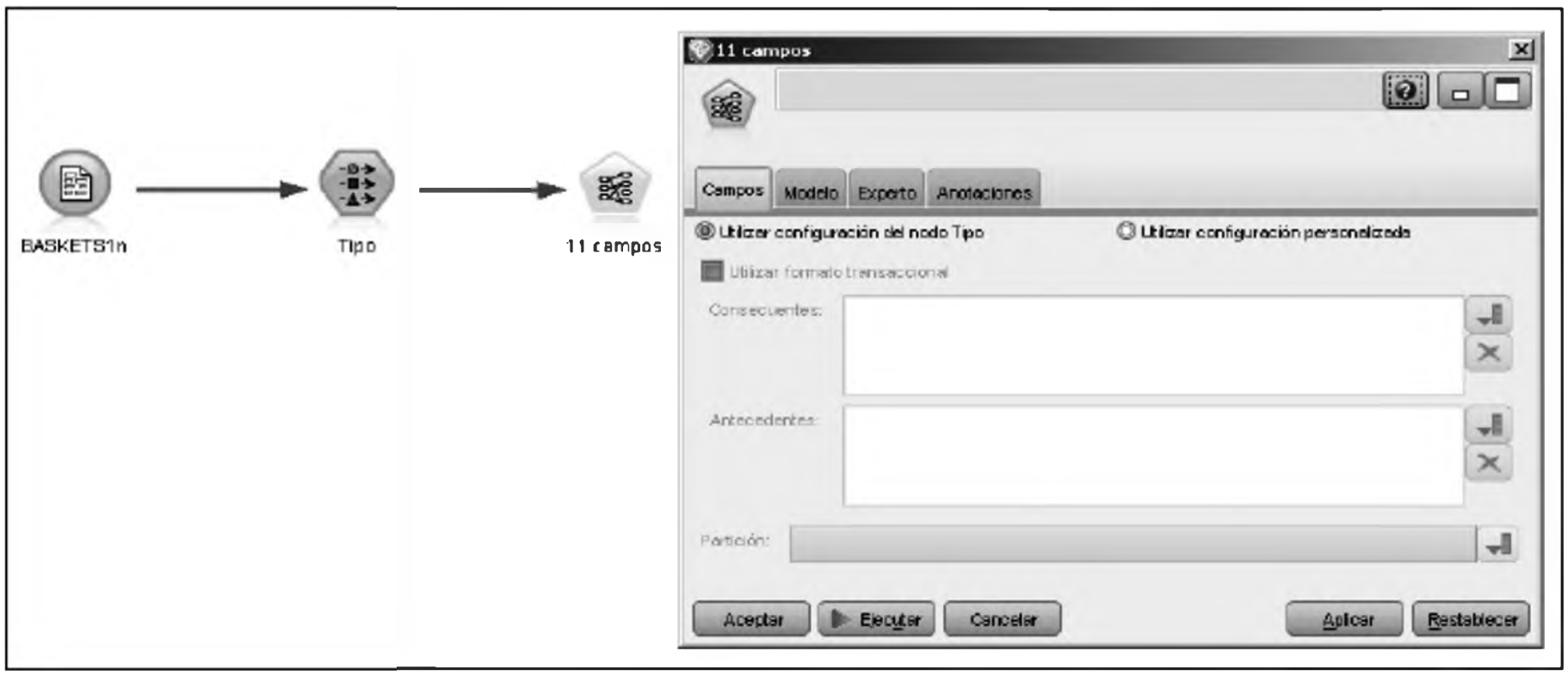


Figura 9-4

Figura 9-5

Figura 9-6

La solapa *Modelo* de la Figura 9-5 tiene los siguientes campos:

Nombre del modelo. Puede generar el nombre del modelo de forma automática basándose en el campo objetivo o de ID (o en el nombre del tipo de modelo si se especifica ningún campo objetivo), o bien especificar un nombre personalizado.

Soporte mínimo de las reglas. Se puede especificar un criterio de soporte para mantener las reglas en el conjunto de reglas. Soporte hace referencia al porcentaje de registros de los datos de entrenamiento en los que los antecedentes (la parte de la regla "si") son verdaderos. (Observe que esta definición de soporte es diferente a la que se utiliza en los nodos CARMA y Secuencia. Consulte el tema Opciones de modelo para el nodo Secuencia si desea obtener más información). Si las reglas obtenidas se aplican a subconjuntos de datos muy pequeños, pruebe a aumentar el valor de este parámetro.

Nota: la definición de soporte para A priori se basa en el número de registros con los antecedentes. Sucede de forma contraria que en los algoritmos CARMA y Secuencia, en los que la definición de soporte se basa en el número de registros con todos los elementos de una regla (es decir, los antecedentes y consecuentes). Los resultados de los modelos de asociación muestran tanto el soporte (antecedente) como las medidas de soporte de reglas.

Confianza mínima de las reglas. También se puede especificar un criterio de confianza. La confianza se basa en los registros por los que los antecedentes de la regla son verdaderos y es el porcentaje de esos mismos registros en los que los consecuentes también son verdaderos. Es decir, es el porcentaje de pronósticos basados en la regla que son correctos. Las reglas con una confianza inferior a la especificada en el criterio de precisión se descartan. Si se obtienen demasiadas reglas, pruebe a aumentar el valor de este parámetro. Si se obtienen muy pocas reglas (o casi ninguna), pruebe a disminuir el valor de este parámetro.

Número máximo de antecedentes. Se puede especificar el número máximo de precondiciones de cualquier regla. Se trata de una forma de limitar la complejidad de las reglas. Si las reglas son demasiado complejas o específicas, pruebe a disminuir el valor de este parámetro. Esta configuración también tiene mucha influencia en el tiempo de entrenamiento. Si el entrenamiento del conjunto de reglas que ha creado se toma demasiado tiempo, pruebe a disminuir el valor de este parámetro.

Solo valores verdaderos para las marcas. Si selecciona esta opción para los datos en formato tabular (tabla de verdad), solo se incluirán los valores verdaderos en las

 ©Alfaomega-RC Libros

reglas resultantes. Esto puede ayudar a que las reglas se entiendan con más facilidad. La opción no se aplica a los datos en formato transaccional. Consulte el tema Datos tabulares frente a datos transaccionales si desea obtener más información.

Optimizar. Seleccione opciones diseñadas para aumentar el rendimiento durante la generación de modelos según sus necesidades específicas.

• Seleccione *Velocidad* para indicar al algoritmo que nunca debe recurrir al volcado en disco para mejorar el rendimiento.

• Seleccione *Memoria* para indicar al algoritmo que utilice el volcado en disco cuando lo considere oportuno en detrimento de la velocidad. Esta opción está seleccionada por defecto. Nota: cuando se ejecuta en modo distribuido, esta configuración puede quedar anulada por las opciones del administrador especificadas en options.cfg. Si desea obtener más información, consulte el Manual del administrador de IBM® SPSS® Modeler Server.

La solapa *Experto* (Figura 9-6) presenta opciones que permiten ajustar el proceso de inducción a los usuarios con conocimientos sobre redes neuronales. Para acceder a estas opciones, active el modo *Experto* en la pestaña *Experto*. Los campos de la solapa *Experto* son los siguientes:

Medida de evaluación. A priori admite cinco métodos de evaluación de reglas potenciales.

• *Confianza de la regla*. El método por defecto utiliza la confianza (o precisión) de la regla para evaluar reglas. Para esta medida, se desactiva el Límite inferior de la medida de evaluación, ya que es redundante con la opción *Confianza mínima de las reglas* de la pestaña *Modelo*. Consulte el tema Opciones de modelo para el nodo A priori si desea obtener más información.

• *Diferencia de confianza*. (También denominada diferencia de confianza mínima absoluta con la previa). Esta medida de evaluación es la diferencia absoluta entre la confianza de la regla y su confianza a priori. Esta opción evita sesgos cuando los resultados no se distribuyen uniformemente. Ayuda a evitar que se conserven reglas "obvias". Establezca el límite inferior de la medida de evaluación con relación a la diferencia mínima de confianza o probabilidad con la que desea que se conserven las reglas.

• *Cociente de confianza*. (También denominada diferencia de cociente de confianza establecida en 1). Esta medida de evaluación es igual a 1 menos el cociente de la confianza de la regla con respecto a la anterior (o si el cociente es superior a uno, su inverso). Al igual que la diferencia de confianza, este método tiene en cuenta las distribuciones que no son homogéneas. Es especialmente apropiado para encontrar reglas que pronostican eventos raros. Establezca el límite inferior de la medida de evaluación en función de la diferencia con la que desea que se conserven las reglas.

• *Diferencia de información*. (También denominada diferencia de información respecto a la previa). Esta medida se basa en la medida de la ganancia de información. Si se considera la probabilidad de un consecuente determinado como un valor lógico (un bit), la ganancia de información es la proporción que puede determinarse de ese bit en función de los antecedentes. La diferencia de información es la existente entre la ganancia de información, dados los antecedentes, y la ganancia de información, dada solo la confianza previa del consecuente. Una característica importante de este método es que tiene en cuenta el soporte de forma que son preferibles aquellas reglas que cubren más registros para un nivel de confianza determinado. Establezca el límite inferior de la medida de evaluación en función de la diferencia de información con la que desea que se conserven las reglas.

• *Chi-cuadrado normalizada*. (También denominada medida de chi-cuadrado normalizada). Esta medida es un índice estadístico de asociación entre antecedentes y consecuentes. La medida se normaliza para adquirir valores entre 0 y 1 y depende aún más del soporte que la medida de la diferencia de información. Establezca el límite inferior de la medida de evaluación en función de la diferencia de información con la que desea que se conserven las reglas.

Permitir reglas sin antecedentes. Seleccione esta opción para permitir las reglas que solo incluyen el consecuente (elemento o conjunto de elementos). Esto resulta de utilidad para determinar elementos o conjuntos de elementos comunes. Por ejemplo, cannedveg es una regla compuesta por un único elemento que carece de antecedentes e indica que la adquisición de cannedveg es una instancia común en los datos. En algunos casos, se pueden incluir estas reglas si solo le interesan los pronósticos de mayor confianza. Esta opción está desactivada por defecto. Por convención, el soporte de antecedentes para las reglas que carecen de antecedentes se expresa con el 100% y el soporte de reglas es el mismo que la confianza.

©Alfaomega-RC Libros

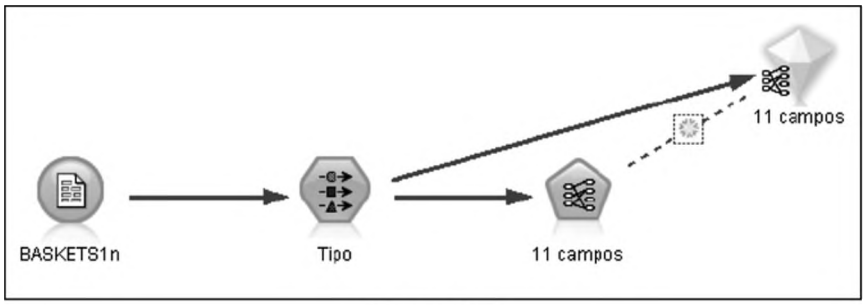


Figura 9-7

Si ahora hecemos clic con el botón derecho del ratón en el nodo *A priori* sobre la ruta, se obtiene la pantalla de resultados de la Figura 9-8. La pestaña *Modelo* muestra una visualización gráfica del modelo de Regla de asociación en la que se incluye una tabla con las reglas que ha extraído el algoritmo. Cada fila de la tabla representa una regla. La primera columna representa los consecuentes (la parte "entonces" de una regla), mientras que la siguiente columna representa los antecedentes (la parte "si" de la regla). Las siguientes columnas contienen información de las reglas, como la confianza, el soporte y la elevación.

Figura 9-8

En las reglas de asociación de la Figura 9-8 (primera fila) se observa que quien consume cerveza y verduras enlatadas también consume comidas congeladas. De igual forma se pueden leer las reglas de asociación de cada fila de la tabla.

REGLAS DE ASOCIACIÓN EN IBM SPSS MODELER. EL NODO CARMA

El nodo CARMA utiliza un algoritmo de descubrimiento de reglas para encontrar reglas de asociación existentes en los datos. Las reglas de asociación son instrucciones del tipo

si antecedente(s) entonces consecuente(s)

Por ejemplo, si un cliente del sitio Web adquiere una tarjeta y un enrutador de gama alta inalámbricos, es muy probable que también adquiera un servidor de música inalámbrico si se le ofrece. El modelo CARMA extrae un conjunto de reglas de los datos sin necesidad de especificar campos de entrada ni de objetivo. Esto significa que las reglas generadas se pueden utilizar en una variedad de aplicaciones mucho más amplia. Por ejemplo, las reglas que ha generado este nodo se pueden utilizar para buscar una lista de productos o servicios (antecedentes) cuyo consecuente es el elemento que desea promocionar durante esta temporada de vacaciones. Con IBM® SPSS® Modeler, puede determinar los clientes que han adquirido los productos antecedentes y diseñar una campaña de marketing destinada a la promoción del producto consecuente.

El nodo CARMA no requiere que los campos sean de Entrada o de Objetivo. Esto es esencial para el modo en que funciona el algoritmo y equivale a la generación de un modelo de A priori con todos los campos establecidos en Ambas. Se pueden restringir los elementos que aparecen solo como antecedentes o como consecuentes activando el filtrado del modelo una vez generado este. Por ejemplo, se puede utilizar el explorador de modelos para buscar una lista de productos o servicios (antecedentes) cuyo consecuente es el elemento que se desea promocionar durante esta temporada de vacaciones.

Para crear un conjunto de reglas de CARMA, es necesario especificar un campo de ID y uno o varios campos de contenido. El campo de ID puede tener cualquier papel o nivel de medición. Los campos con el papel *Ninguno* se omiten. Los tipos de campo deben estar completamente instanciados antes de ejecutar el nodo. Al igual que en A priori, los datos pueden estar en formato tabular o transaccional.

El nodo CARMA se basa en el algoritmo de reglas de asociación de CARMA. A diferencia de A priori, el nodo CARMA ofrece opciones de construcción basadas en el soporte de la regla (soporte tanto para el antecedente como el consecuente) en lugar de hacerlo solo respecto al soporte del antecedente. CARMA también permite reglas con varios consecuentes. Como sucede con A priori, los modelos que genera un nodo CARMA se pueden insertar en una ruta de datos para crear pronósticos.

 ©Alfaomega-RC Libros

Al igual que en el caso del nodo anterior Ilustramos el algoritmo con un ejemplo que busca las asociaciones entre los productos alimenticios de la dieta de jugadores de baloncesto. Las variables se recopilan en el archivo ASCII *baskets1n*.

Comenzamos utilizando como origen de datos el conjunto de datos ASCII en formato variable *baskets1n.txt* y lo unimos con un nodo *Tipo* para situar como variables de medida *Marca* y papel *Ambas* todas las involucradas en el análisis de asociaciones (variables de tipo alimenticio) y como no utilizables el resto de las variables del archivo salvo la variable *cardid* con papel *ID de registro* (Figura 9-9).

Para realizar el análisis de asociaciones añadimos el nodo *Carma* de la categoría *Asociación* de la ficha *Modelado* al diagrama. A continuación hacemos clic con el botón derecho del ratón sobre el nodo *Carma* y elegimos *Edición* en el menú emergente resultante. Se obtiene la pantalla de la Figura 9-10 en cuya solapa *Campos* elegimos *Utilizar la configuración del nodo tipo*. En la solapa *Modelo* elegimos *Automático y Utilizar los datos en particiones* (Figura 9-11). En al solapa *Experto* (Figura 9-12) elegimos *Experto* y el resto de opciones por defecto. Al hacer clic en *Aplicar* y *Ejecutar* se obtiene el nodo *Carma* sobre la ruta (Figura 9-12).

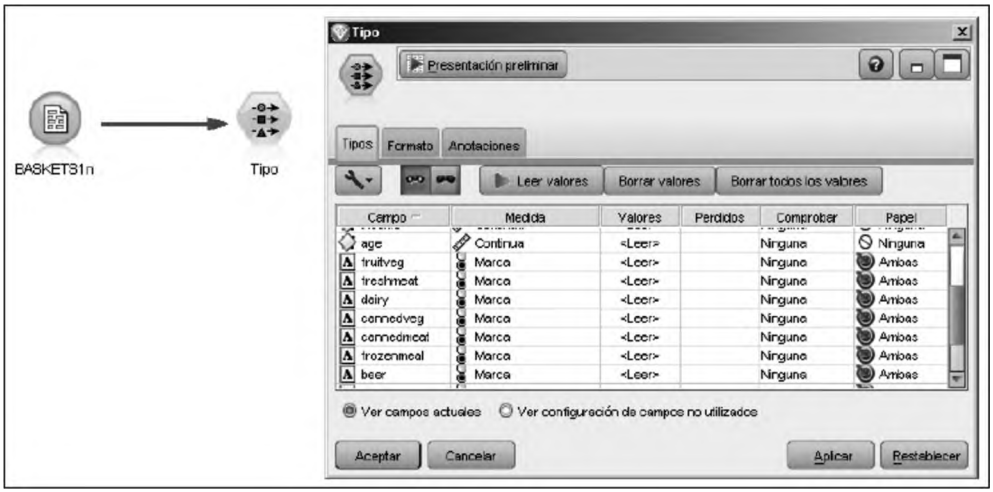


Figura 9-9

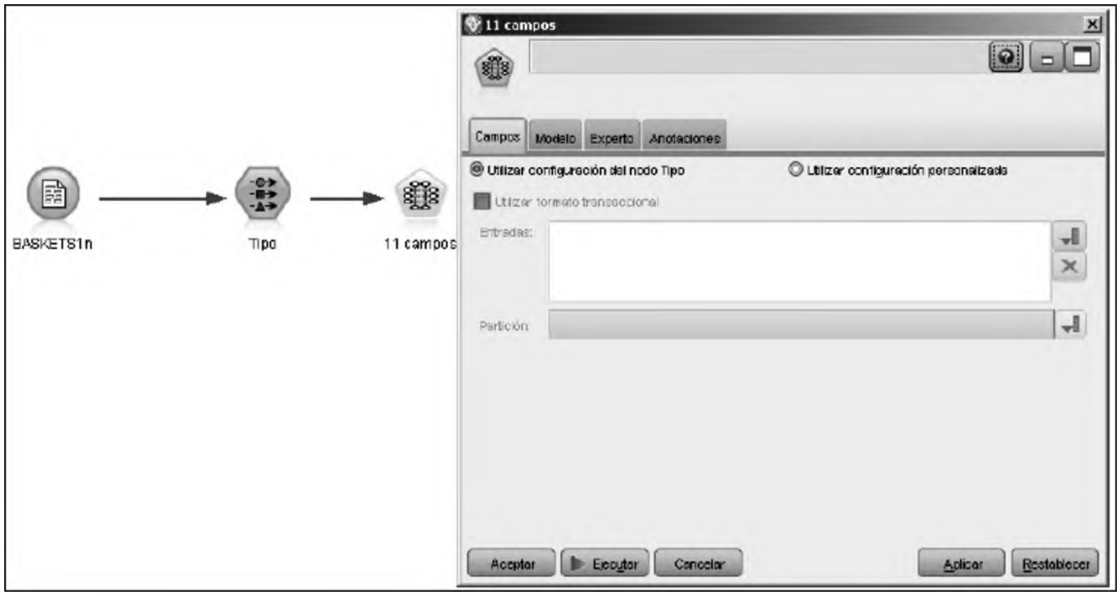


Figura 9-10

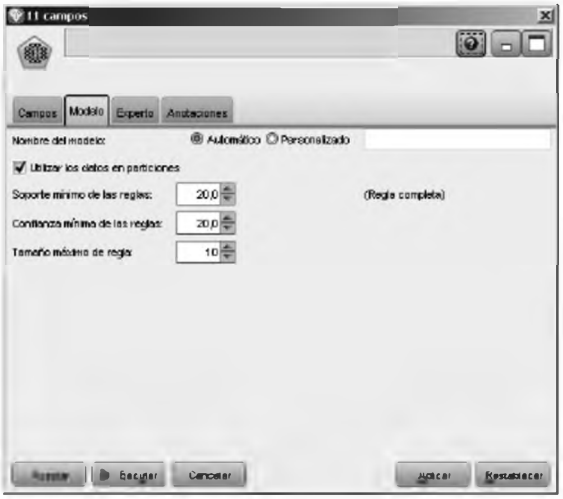


Figura 9-11

Figura 9-12

Antes de ejecutar un nodo CARMA se deben especificar los **campos de entrada en la pestaña Campos del nodo CARMA** (Figura 9-10). Mientras que la mayoría de los nodos de Modelado comparten las mismas opciones de la pestaña *Campos*, el nodo CARMA contiene muchas opciones particulares. A continuación se describen todas las opciones.

Utilizar configuración del nodo Tipo. Esta opción permite indicar al nodo que use la información de campo de un nodo Tipo situado en un punto anterior de la ruta. Este es el método por defecto.

 ©Alfaomega-RC Libros

Utilizar configuración personalizada. Esta opción permite indicar al nodo que use la información de campo especificada aquí en lugar de la proporcionada en nodos Tipo situados en cualquier punto anterior de la ruta. Una vez seleccionada esta opción, especifique los campos en función del formato (transaccional o tabular) en el que desee leer los datos.

Utilizar formato transaccional. Esta opción modifica los controles de campo del resto de este cuadro de diálogo en función de que el formato de los datos sea transaccional o tabular. Si se utilizan varios campos con datos transaccionales, se asume que los elementos especificados en estos campos para un registro determinado representan los elementos encontrados en una sola transacción con una sola marca de tiempo.

Datos tabulares

Si no se selecciona *Utilizar formato transaccional*, se muestran los siguientes campos:

• *Entradas*. Seleccione el campo(s) de entrada. Se trata de una acción similar a establecer el papel del campo en Entrada en un nodo Tipo.

• *Partición*. Este campo permite especificar un campo usado para dividir los datos en muestras independientes para las fases de entrenamiento, prueba y validación en la generación del modelo. Si usa una muestra para generar el modelo y otra muestra distinta para comprobarlo, podrá obtener una buena indicación sobre la adecuación del modelo a la hora de generar conjuntos de datos de mayor tamaño similares a los datos actuales. Si se han definido varios campos de partición mediante nodos Tipo o Partición, se deberá seleccionar un campo de partición simple en la ficha Campos en todos los nodos de modelado que usen la partición. (Si solo hay una partición, se usará automáticamente siempre que se active la partición). Consulte el tema Nodo Partición si desea obtener más información. Debe tener en cuenta que al aplicar la partición seleccionada en su análisis, también debe activar la partición en la ficha *Opciones* del modelo para el nodo. (Si se elimina la selección de esta opción, se posibilita la desactivación de la partición sin cambiar la configuración del campo).

Datos transaccionales

Si selecciona *Utilizar formato transaccional*, se muestran los siguientes campos.

• ID. Para los datos transaccionales, seleccione el campo de ID de la lista. Los campos numéricos o simbólicos se pueden utilizar como campo de ID. Cada valor único de este campo debe indicar una unidad de análisis específica. Por ejemplo, en una aplicación de la cesta del supermercado, cada ID puede representar a un solo cliente. Para una aplicación de análisis del registro Web, cada ID puede representar un equipo (con la dirección IP) o un usuario (con los datos de inicio de sesión).

• *Los ID son contiguos*. (Nodos Apriori y CARMA únicamente). Si los datos se han clasificado previamente de forma que todos los registros con el mismo ID se agrupan en la ruta de datos, seleccione esta opción para que el procesamiento sea más rápido. Si los datos no se han clasificado previamente (o no lo sabe a ciencia cierta), no active esta opción y el nodo clasificará los datos automáticamente.

Nota: si los datos no están clasificados y selecciona esta opción, es posible que obtenga resultados no válidos en el modelo.

• *Contenido*. Especifique los campos de contenido del modelo. Estos campos contienen los elementos de interés del modelo de asociación. Se pueden especificar varios campos de marcas (si los datos están en formato tabular) o un solo campo nominal (si los datos están en formato transaccional).

En la Figura 9-11 se observan las **opciones de modelo para el nodo CARMA**. Son las siguientes:

Nombre del modelo. Puede generar el nombre del modelo de forma automática basándose en el campo objetivo o de ID (o en el nombre del tipo de modelo si se especifica ningún campo objetivo), o bien especificar un nombre personalizado.

Soporte mínimo de las reglas (%). Puede especificar un criterio de soporte. Soporte de la regla hace referencia a la proporción de campos de ID existente en los datos de entrenamiento que contienen la regla completa. (Tenga en cuenta que esta definición de soporte es diferente al soporte del antecedente utilizado en los nodos A priori). Si desea centrarse en las reglas más comunes, aumente el valor de este parámetro.

Confianza mínima de las reglas (%). Se puede especificar un criterio de confianza para mantener las reglas en el conjunto de reglas. La confianza hace referencia al porcentaje de campos de ID en los que se realiza un pronóstico correcto (de todos los campos de ID para los que la regla realiza un pronóstico). Se calcula como la cantidad de ID en los que se encuentra la regla completa dividido por la cantidad de ID en los que se encuentran los antecedentes, basado en los datos de entrenamiento. Las reglas con una confianza inferior a la especificada en el criterio de precisión se

 ©Alfaomega-RC Libros

descartan. Si se obtienen demasiadas reglas o reglas de poco interés, pruebe a aumentar el valor de este parámetro. Si se obtienen muy pocas reglas, pruebe a disminuir el valor de este parámetro.

Tamaño máximo de regla. Se puede configurar el número máximo de conjuntos de elementos (a diferencia de los elementos) distintos en una regla. Si las reglas de interés resultantes son pocas, se puede disminuir el valor del parámetro para que el conjunto de reglas se genere más rápido.

En la Figura 9-12 se observan las *opciones de experto para el nodo CARMA.* Las opciones de experto siguientes permiten ajustar el proceso de generación de modelos a los usuarios con conocimientos sobre el funcionamiento del nodo CARMA. Para acceder a estas opciones, active el modo *Experto* en la pestaña *Experto*.

Excluir reglas con varios consecuentes. Seleccione esta opción para excluir los "consecuentes "de dos destinos, es decir, los que contienen dos elementos. Por ejemplo, la regla bread & cheese & fish -> wine&fruit contiene una regla consecuente de dos direcciones, wine&fruit. Estas reglas se incluyen por defecto.

Definir valor de poda. Para conservar la memoria, el algoritmo CARMA utilizado periódicamente elimina (poda) los conjuntos de elementos poco frecuentes de una lista de conjuntos de elementos potenciales durante el procesamiento. Seleccione esta opción para ajustar la frecuencia de poda; el número especificado determina la frecuencia de la misma. Introduzca un valor más pequeño para disminuir los requisitos de memoria del algoritmo (pero aumentar potencialmente el tiempo de entrenamiento necesario) o introduzca un valor mayor para que el entrenamiento sea más rápido (pero aumentar potencialmente los requisitos de memoria). El valor por defecto es 500.

Variar soporte. Seleccione esta opción para aumentar la eficacia mediante la exclusión de conjuntos de elementos poco frecuentes que aparentan ser frecuentes cuando se incluyen de forma irregular. Esto se consigue comenzando con un nivel de soporte superior que después se disminuye hasta el nivel especificado en la pestaña *Modelo*. Especifique un valor en Número de transacciones estimado para especificar la velocidad con la que debe disminuirse el nivel de soporte.

Permitir reglas sin antecedentes. Seleccione esta opción para permitir las reglas que solo incluyen el consecuente (elemento o conjunto de elementos). Esto resulta de utilidad para determinar elementos o conjuntos de elementos comunes. Por ejemplo, cannedveg es una regla compuesta por un único elemento que carece de antecedentes e indica que la adquisición de cannedveg es una instancia común en los

datos. En algunos casos, se pueden incluir estas reglas si solo le interesan los pronósticos de mayor confianza. Esta opción está desactivada por defecto.

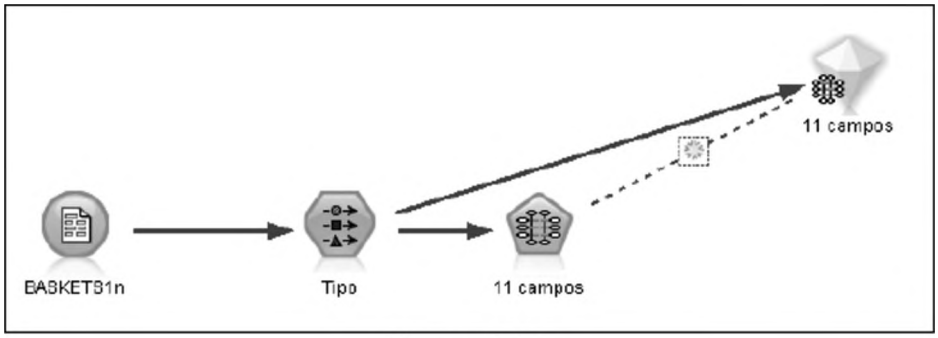
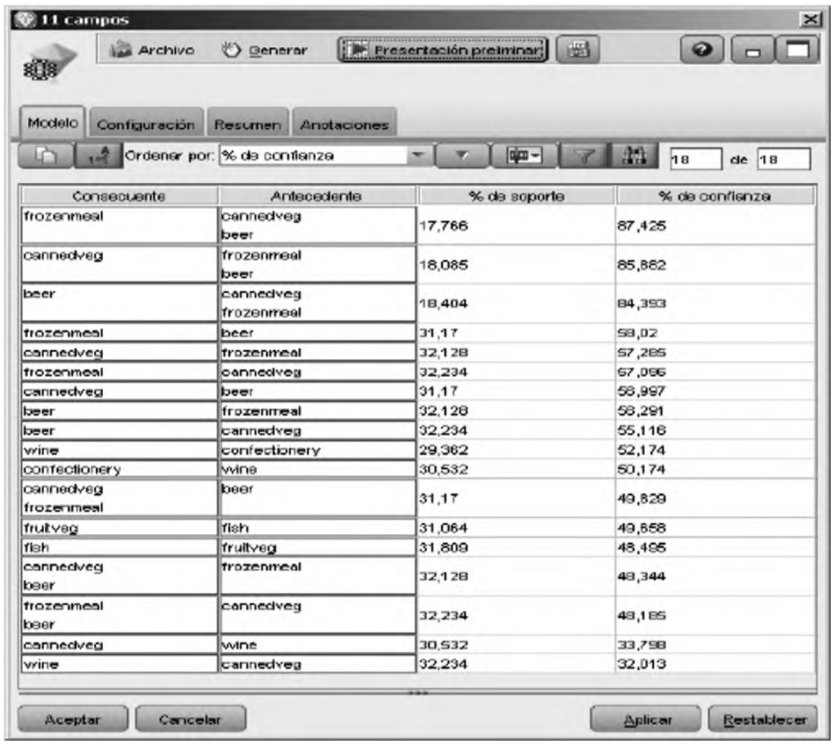


Figura 9-13

Si ahora hecemos clic con el botón derecho del ratón en el nodo *A priori* sobre la ruta, se obtiene la pantalla de resultados de la Figura 9-14. La pestaña *Modelo* muestra una visualización gráfica del modelo de Regla de asociación en la que se incluye una tabla con las reglas que ha extraído el algoritmo. Cada fila de la tabla representa una regla. La primera columna representa los consecuentes (la parte "entonces" de una regla), mientras que la siguiente columna representa los antecedentes (la parte "si" de la regla). Las siguientes columnas contienen información de las reglas, como la confianza, el soporte y la elevación.



Consecuente	Antecedente	% de soporte	% de confianza
frozenmeal	cannedveg beer	17,766	87,425
cannedveg	frozenmeal beer	18,085	85,862
beer	cannedveg frozenmeal	18,404	84,353
frozenmeal	beer	31,17	58,02
cannedveg	frozenmeal	32,128	57,285
frozenmeal	cannedveg	32,234	57,096
cannedveg	beer	31,17	56,997
beer	frozenmeal	32,128	56,291
beer	cannedveg	32,234	55,116
wine	confectionery	29,362	52,174
confectionery	wine	30,532	50,174
cannedveg frozenmeal	beer	31,17	49,829
fruitveg	fish	31,064	49,658
fish	fruitveg	31,809	48,495
cannedveg beer	frozenmeal	32,128	48,344
frozenmeal beer	cannedveg	32,234	48,185
cannedveg	wine	30,532	33,798
wine	cannedveg	32,234	32,013

Figura 9-14

©Alfaomega-RC Libros

En las reglas de asociación de la Figura 9-14 (primera fila) se observa que quien consume cerveza y verduras enlatadas también consume comidas congeladas. De igual forma se pueden leer las reglas de asociación de cada fila de la tabla.

NODO DE ASOCIACIONES EN SAS ENTERPRISE MINER

El nodo de asociaciones, cuyo icono en SAS Enterprise Miner se presenta en la Figura 9-15, permite identificar elementos o variables que suceden simultáneamente. Esta técnica también es conocida como "análisis de la cesta de la compra". La extracción de reglas de asociación está basada en contar las frecuencias con las que determinados valores aparecen solos o en compañía de otros. Los resultados se expresan como reglas del tipo "si el caso A es parte de un evento entonces el caso B es también parte de ese evento en un porcentaje de veces".

Figura 9-15

Estas reglas no deben ser interpretadas como causas directas sino como fenómenos que suelen aparecer asociados. En las reglas de asociación tampoco importa si por ejemplo el cliente compra una o varias unidades de A, solo cuenta la presencia de A en la cesta de la compra. Sin embargo, en marketing estas asociaciones son de suma utilidad para por ejemplo colocar los productos en las estanterías, repartir cupones de descuento, decidir cuándo sacar un producto a la venta, ofrecer descuentos, etc.

Algunos ejemplos de asociaciones hipotéticas serían las siguientes:

- Si un cliente compra un refresco de cola entonces el 30% de las veces también compra algún tipo de aperitivo o snack.
- Si un cliente compra cuchillas de afeitado en un 20% de ocasiones también compra espuma de afeitado.
- Si un cliente compra comida congelada y pañales en un 15% de ocasiones también compra latas de cerveza.

Estas reglas tienen un antecedente (la compra de un refresco) y una consecuencia (la compra de un snack). Ambos lados de la regla de asociación pueden

contener más de un elemento. La bondad de una regla de asociación se mide de acuerdo a una serie de criterios:

- ***Factor de confianza (Confidence factor):*** mide la fuerza de la asociación medida como el porcentaje de casos en los cuales la consecuencia aparece una vez que el antecedente se ha producido.

- ***Nivel de seguimiento (Level of support):*** mide el nivel de frecuencia en el que las combinaciones ocurren en la base de datos.

- ***Confianza esperada (Expected confidence):*** número de apariciones relevantes sobre el total de transacciones.

- ***Factor de elevación o de ganancias* (*Lift*)**: Se define como el factor de confianza dividido por la confianza esperada. Es una medida de las ganancias o probabilidades de suceso que pueden obtenerse aplicando la regla.

A partir de la primera regla puesta de ejemplo en el caso anterior observamos los siguientes resultados en la cadena de supermercados de donde la información fue tomada:

- Ventas de Unidades: 2000000
- Refrescos de cola vendidos: 50000
- Snacks vendidos: 35000
- Refrescos y Snacks vendidos simultáneamente: 15000

Factor de confianza: 15000/50000 = 30%
Seguimiento (Support): 15000/2000000 = 0.75%
Confianza esperada: 35000/2000000 = 1.75%
Lift (ganancia) = Factor de confianza / Confianza Esperada = 17

En este caso el Lift mostraría que es 17 veces más probable comprar un snack una vez se compra el refresco de cola que si este no se compra.

Para encontrar este tipo de asociaciones los datos deben estar dispuestos de tal manera que cada producto a evaluar debe tener una observación. Si los datos no están ordenados de esta forma se debe escribir un comando PROC DATA o PROC TRANSPOSE en el nodo de código SAS para crear un nuevo fichero de datos que contenga una observación separada de cada producto comprado por cada cliente.

 ©Alfaomega-RC Libros

Además el nodo necesita que al menos una variable tenga un rol ID y otra un rol de objetivo categórico.

Dentro del nodo de búsqueda de asociaciones se pueden establecer las siguientes diferentes opciones que aparecen en la línea superior de la Figura 9-16.

Opción General

La Figura 9-16 muestra los campos de la opción *General*, que se comentarán a continuación.

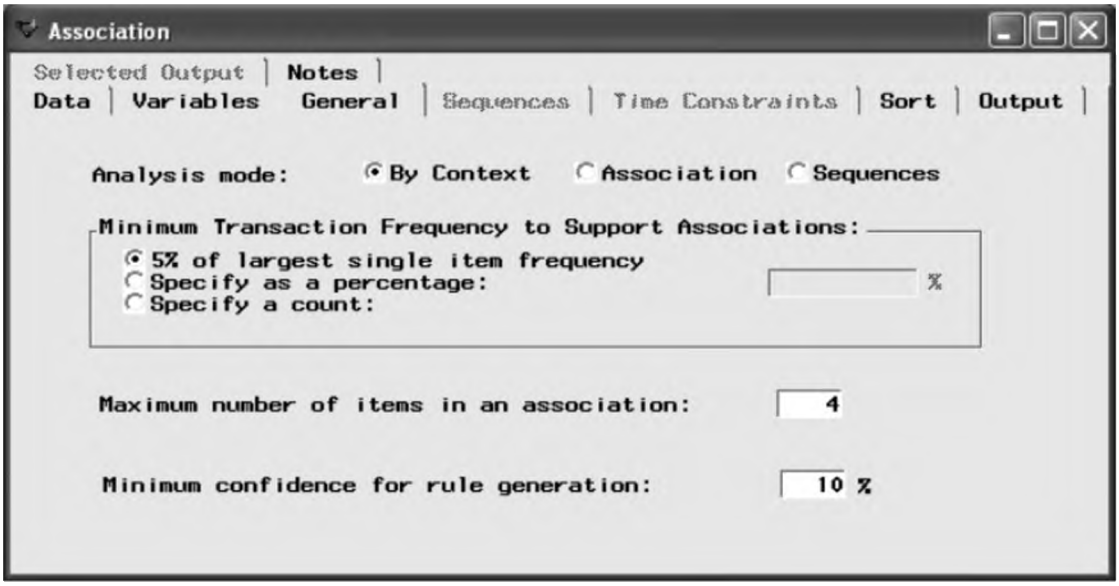


Figura 9-16

- ***Modo de análisis (Analysis mode):*** la opción *By Context* asigna automáticamente el tipo de análisis. Si los datos contienen una variable de identidad (*id*) y una variable objetivo (*target*) entonces el sistema automáticamente lleva a cabo un análisis de asociación. Si los datos también contienen una variable de secuencia con un status de "en uso" entonces el sistema llevará a cabo un análisis de secuencia.

- ***Frecuencia mínima para considerar asociaciones (Minimum Transaction Frequency to Support Associations):*** en esta opción es necesario establecer un nivel mínimo para que la regla de asociación sea generada. Por defecto se asigna una frecuencia del 5% si bien este porcentaje puede ser cambiado por otro o por un número mínimo de casos.

- ***Número máximos de elementos asociados (Maximum number of items in an association):*** determina el tamaño máximo de elementos que pueden estar asociados.

- ***Confianza mínima para generar una regla (Minimum confidence for rule generation):*** por defecto se establece en el 10% aunque puede ser modificada. Ello es especialmente aconsejable para generar pocas reglas aunque con un elevado nivel de confianza.

Opción Sequence

Para el descubrimiento de secuencias disponemos de los campos de la Figura 9-17 y que se describen a continuación.

- ***Minimum Transaction Frequency to Support Sequence:*** permite filtrar aquellas secuencias que ocurren infrecuentemente, por defecto es menos del 2%. Se puede especificar otro valor tanto en porcentaje como en número de veces.

- ***Number of items in longest chain:*** permite establecer el número máximo de elementos a incluir en una secuencia, por defecto es 3. Para cambiarlo simplemente introduzca un nuevo valor que en ningún caso debe ser mayor que 10.

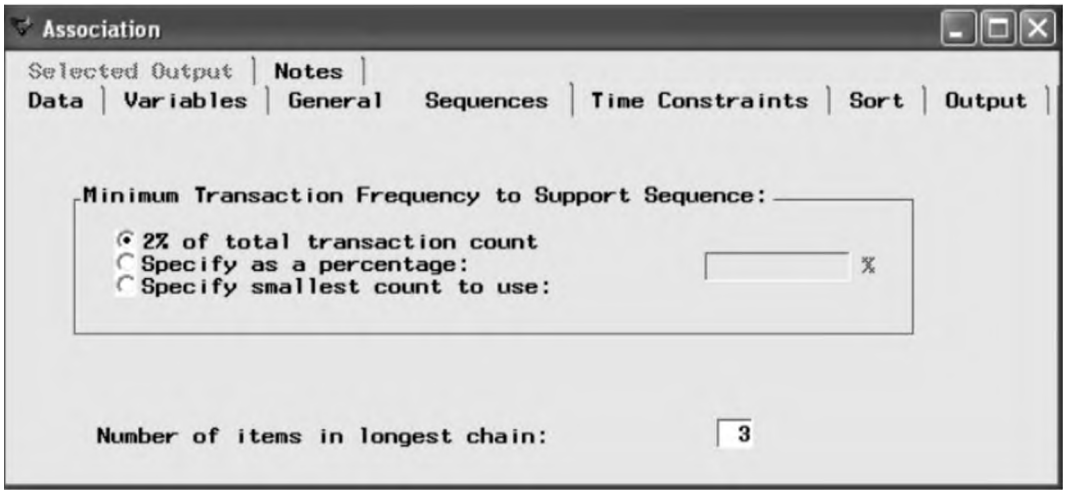


Figura 9-17

©Alfaomega-RC Libros

Opción Time Constraints

En esta pestaña (Figura 9-18) es posible establecer la longitud de la transacción para descubrir secuencias. La unidad de medida será determinada por los datos aunque debe estar definida en un único rango (segundos, minutos, horas, etc.).

- ***Transaction Window Length:*** por defecto se identifican todas las secuencias. Si queremos limitar el tiempo deberemos activar Specify duration to use y definir el valor relevante para el análisis.

- ***Consolidate time differences:*** esta opción permite consolidar distintos momentos temporales en uno solo. Por ejemplo si un cliente acude tres veces al supermercado podemos consolidar estas tres veces como si acudiera a comprar una sola vez en el mismo día.

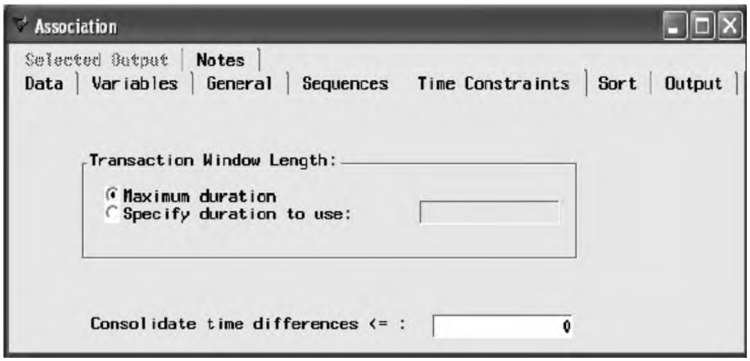


Figura 9-18

Opción Sort

Esta ventana (Figura 9-19) permite ordenar las variables de etiqueta en el caso de que existan varias. Para ello estas deben ser incluidas en la lista *Selected* en el orden deseado. Este orden determina la interpretación de las asociaciones. Por ejemplo supongamos que tuviéramos dos variables de etiqueta: TIENDA y CLIENTE. El orden STORE (Primaria) – CLIENTE (Secundaria) explora los hábitos de compra de los clientes dentro de cada tienda. Por el contrario el orden CLIENTE (Primaria) – TIENDA (Secundaria) exploraría las ventas de cada tienda para cada cliente.

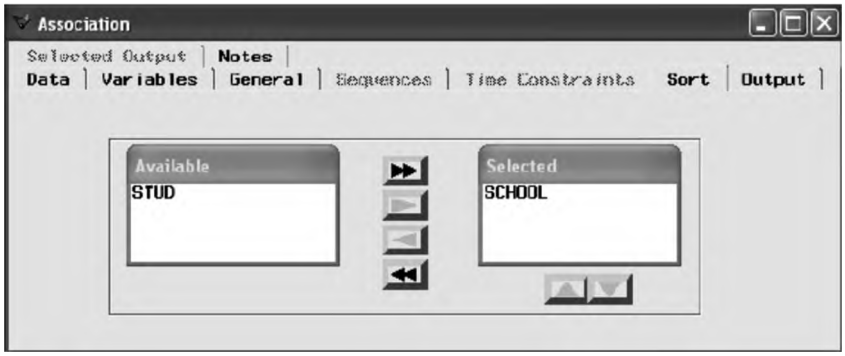


Figura 9-19

Una vez ejecutado el nodo podremos visualizar los resultados para cada una de las reglas generadas (Figura 9-20). En esta pantalla se muestran las asociaciones entre el nivel de estudios de los padres de los alumnos de la base de datos BÉLGICA. Para ello la variable de etiqueta (ID) es la escuela, el objetivo los estudios de la madre y el input el nivel de estudios del padre.

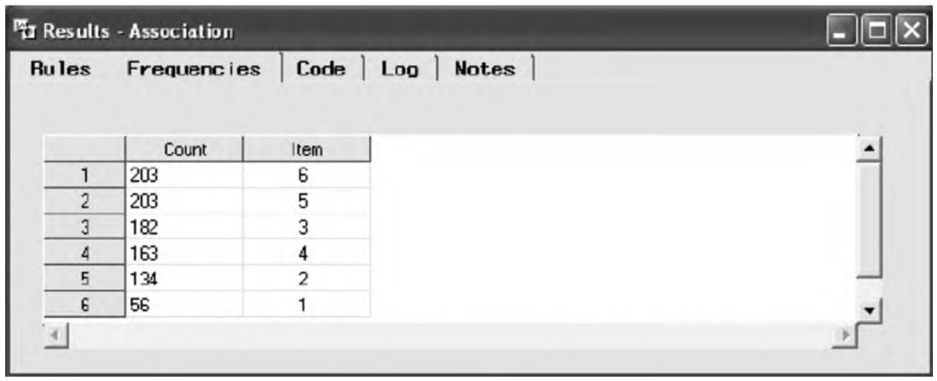
Rules | Frequencies | Code | Log | Notes

	Lift	Support(%)	Confidence(%)	Transaction Count	Rule
1	1.01	91.55	96.06	195.00	6 ==> 5
2	1.01	91.55	96.06	195.00	5 ==> 6
3	1.01	82.63	86.70	176.00	5 ==> 3
4	1.01	82.63	96.70	176.00	3 ==> 5
5	1.01	82.16	86.21	175.00	6 ==> 3
6	1.01	82.16	96.15	175.00	3 ==> 6
7	1.02	74.18	77.83	158.00	6 ==> 4
8	1.02	74.18	96.93	158.00	4 ==> 6
9	1.02	74.18	77.83	158.00	5 ==> 4
10	1.02	74.18	96.93	158.00	4 ==> 5
11	1.03	67.14	87.73	143.00	4 ==> 3
12	1.03	67.14	78.57	143.00	3 ==> 4

Figura 9-20

La pestaña *Frequencies* muestra la distribución de la frecuencia de las unidades (escuelas) que presentan el valor objetivo al menos una vez (Figura 9-21). Es por ello que estos valores no coinciden con la verdadera distribución en la variable objetivo.

©Alfaomega-RC Libros



	Count	Item
1	203	6
2	203	5
3	182	3
4	163	4
5	134	2
6	56	1

Figura 9-21

El editor de resultados también puede mostrar gráficamente las asociaciones encontradas. Previamente en la tabla mostrada en la pestaña *Rules* debemos seleccionar aquellas reglas que queramos representar. Una vez seleccionadas, en el menú principal seleccionaremos *View* →*Graph* y se obtiene el gráfico de la Figura 9-22.

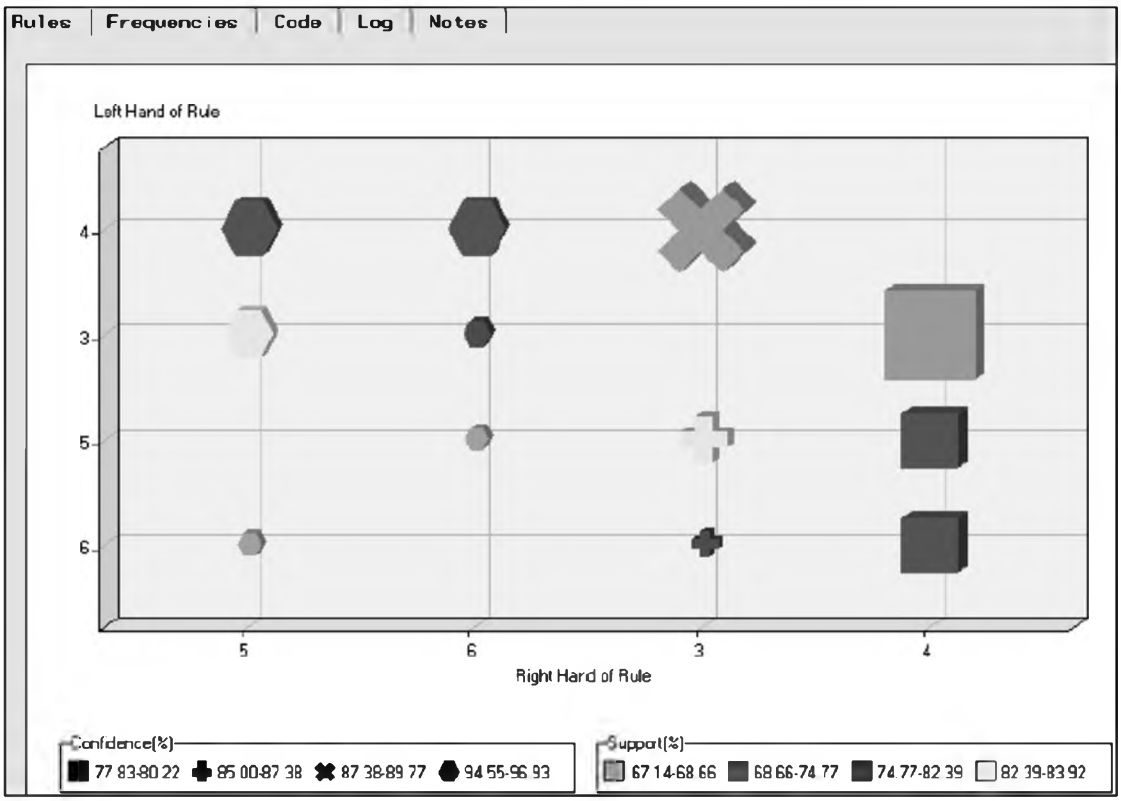


Figura 9-22

NODO DE ANÁLISIS DE RELACIONES EN SAS ENTERPRISE MINER

El análisis de relaciones, cuyo icono en SAS Enterprise Miner se presenta en la Figura 9-23), consiste en examinar los vínculos entre distintos efectos dentro de un sistema más o menos complejo. Las técnicas para este tipo de análisis incluyen los cubos OLAP (*On-Line Analytic Processing*), asociaciones, secuencias, cluster y análisis gráfico.

Figura 9-23

Las aplicaciones de esta utilidad son la detección del fraude, visualización de bases de datos y el análisis de relaciones sociales.

El nodo permite distintos métodos de entrenamiento. Todos ellos devuelven dos ficheros de datos de resultados:

- **NODOS**: el fichero de nodos contiene una fila para cada nodo de un gráfico de relaciones. Cada nodo contiene una única variable de etiqueta (ID) y varias columnas con información descriptiva.

- **RELACIONES** (LINKS): este fichero contiene una fila para cada relación de un gráfico de relaciones. Cada fila debe contener dos variables de etiqueta además de columnas con información descriptiva.

Los gráficos de este nodo muestran las relaciones entre los niveles de las variables usando la información de los nodos y las relaciones. El nodo (Figura 9-24), además de las pestañas habituales presenta un menú de opciones avanzadas. Para acceder a ellas seleccionaremos en la barra de herramientas el icono .

©Alfaomega-RC Libros

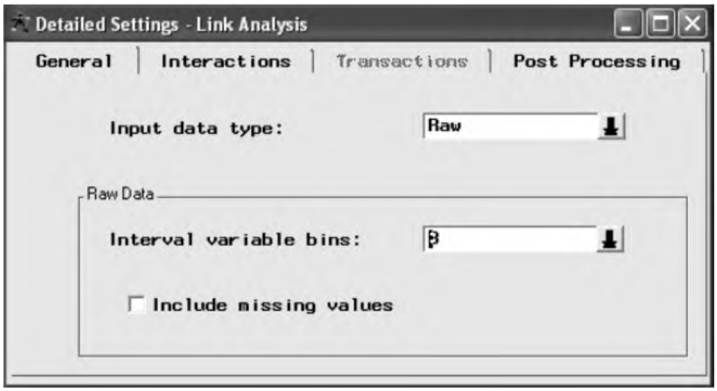


Figura 9-24

Opción General

Permite configurar el tipo de datos disponibles para el análisis:

- ***Raw Data***: es la forma más frecuente en *Enterprise Miner*. El comando PROC FREQ construye los datos con los nodos univariables y los datos con las uniones bivariantes.
- ***Matrix Data***: se utiliza cuando los datos están en forma de matriz de distancias. Los datos de nodos y vínculos serán derivados de esta matriz.
- ***Transaction Data***: los datos de análisis deben contener variables de etiqueta y objetivo. El comando PROC ASSOC crea los datos con los nodos y vínculos. Los valores indivuduales de las variables objetivo son los nodos del gráfico.
- ***Transaction data with time element***: en esta opción los datos deben contener una variable de etiqueta, una variable objetivo y una de secuencia.

La opción *Interval variable bins* especifica el número de categorías que queremos usar para las variables continuas.

Opción Interactions

Permite configurar que interacciones queremos incluir y excluir del análisis. Si tenemos muchos inputs las interacciones pueden hacer que los resultados sean prácticamente imposibles de interpretar. Las interacciones a excluir serán llevadas a la ventana *Exclude* (Figura 9-25).

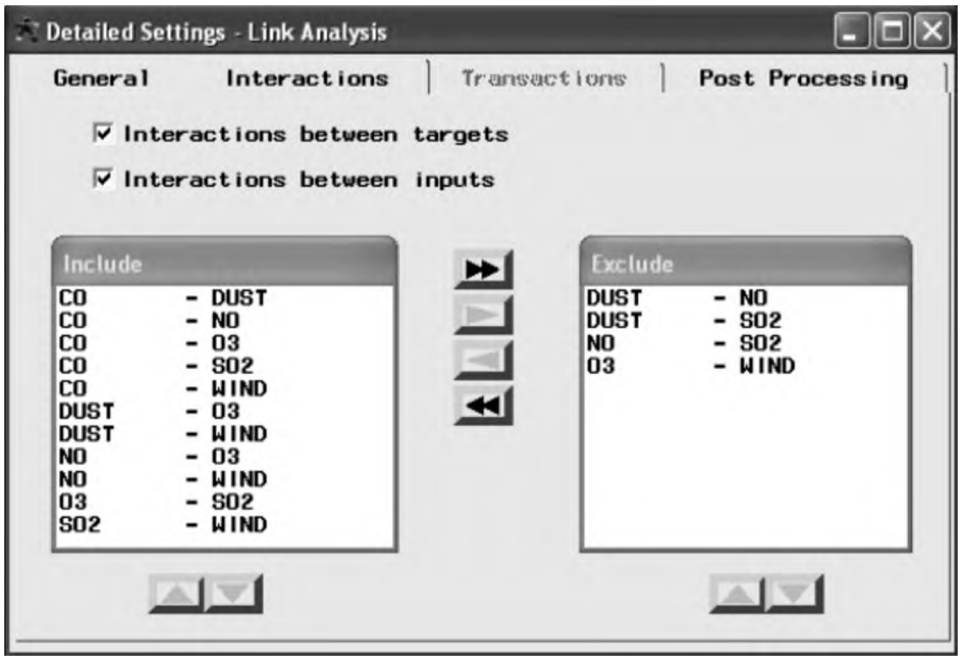


Figura 9-25

Opción Transactions

Si no existen variables de secuencia o etiqueta está ventana no estará disponible. Si disponemos de estas variables podremos definir el umbral mínimo (*Minimum support*) necesario para definir una asociación como significativa así como el número máximo de elementos que permitiremos estén presentes en una asociación (*Max. number of items*) tal y como se indica en la Figura 9-26.

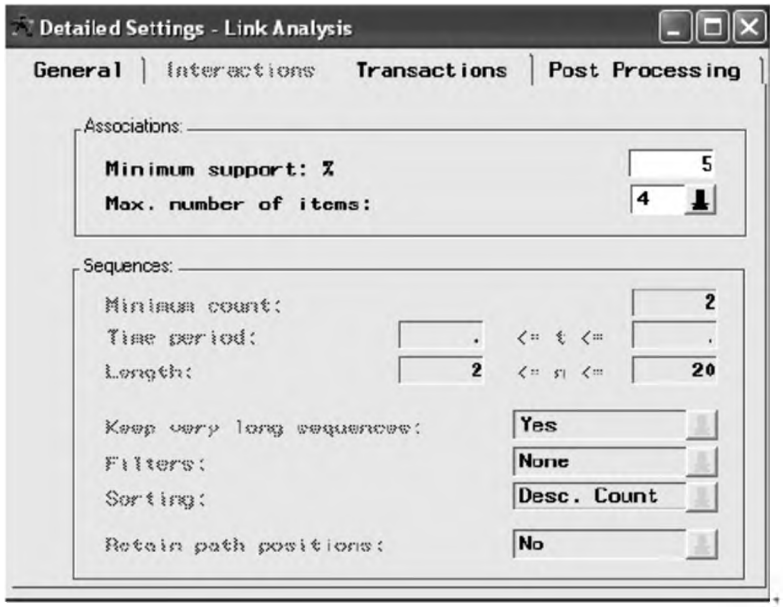


Figura 9-26

©Alfaomega-RC Libros

Si tenemos variables de secuencia se activarán las opciones correspondientes para configurar cómo serán las secuencias definidas y usadas en el análisis (Figura 9-27):

- ***Minimum Count***: número mínimo de elementos contados que definen una secuencia.
- ***Time period***: rango de tiempo permitido en una secuencia.
- ***Length:*** rango de números permitidos dentro de una secuencia.
- ***Keep very long sequences***: permitir secuencias largas (a veces difíciles de interpretar).
- ***Filters***: configura cómo crear secuencias (Repetición, Duplicar).
- ***Sorting***: configura secuencias por orden (Cuenta, cuenta descendente, los primeros N elementos, usando el primer elemento, usando el último elemento).
- ***Retain path positions***: permite transformar secuencias en nodos y relaciones.

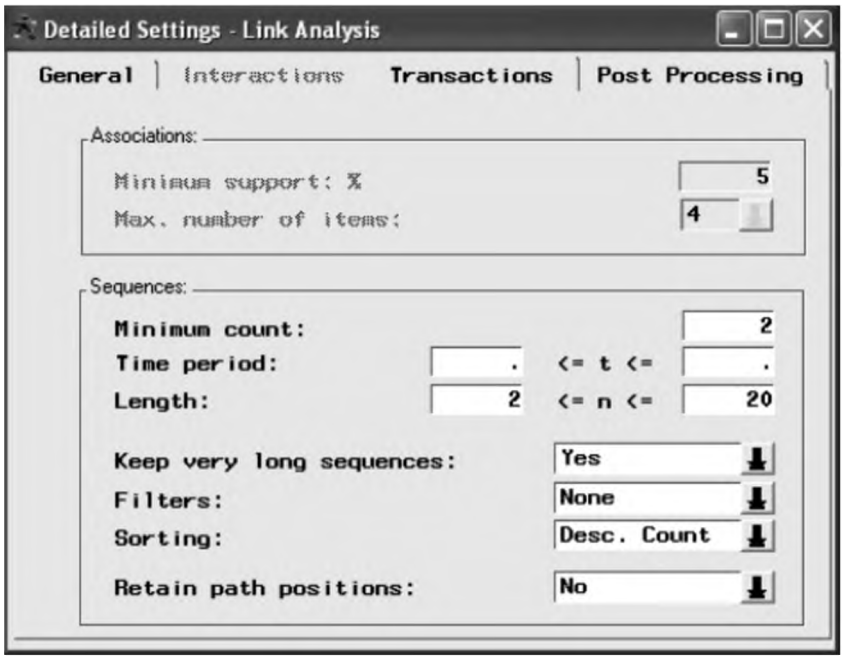


Figura 9-27

Opción Post Processing

Una vez que los nodos y relaciones son creados se pueden configurar distintas opciones para su procesado (Figura 9-28).

- ***Text delimiter:*** delimitador de texto.
- ***Number of delimiters:*** número de delimitadores de texto.
- ***Centrality measures:*** mide la importancia de un nodo dentro del diagrama. La opción *Weighted* asigna un valor entre 0 y 1 para cada nodo según su importancia mientras que *Unweighted* da la misma importancia a cada nodo.
- ***Sort by count order:*** ordenación de los datos output.
- ***Max. Rows saved:*** controla la longitud de los datos output.
- ***Output Matrix:*** permite guardar los datos de resultados en forma de matriz para ser usada posteriormente.

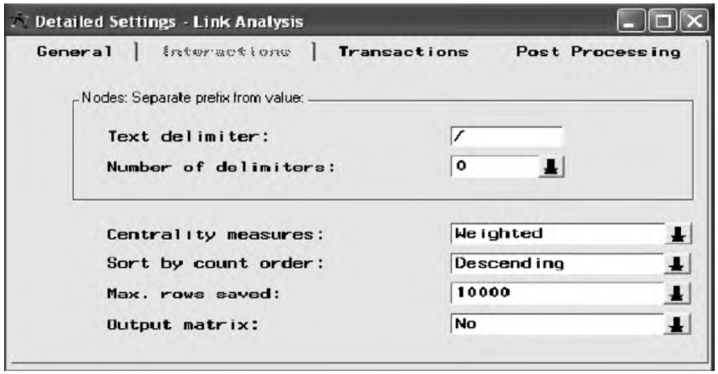
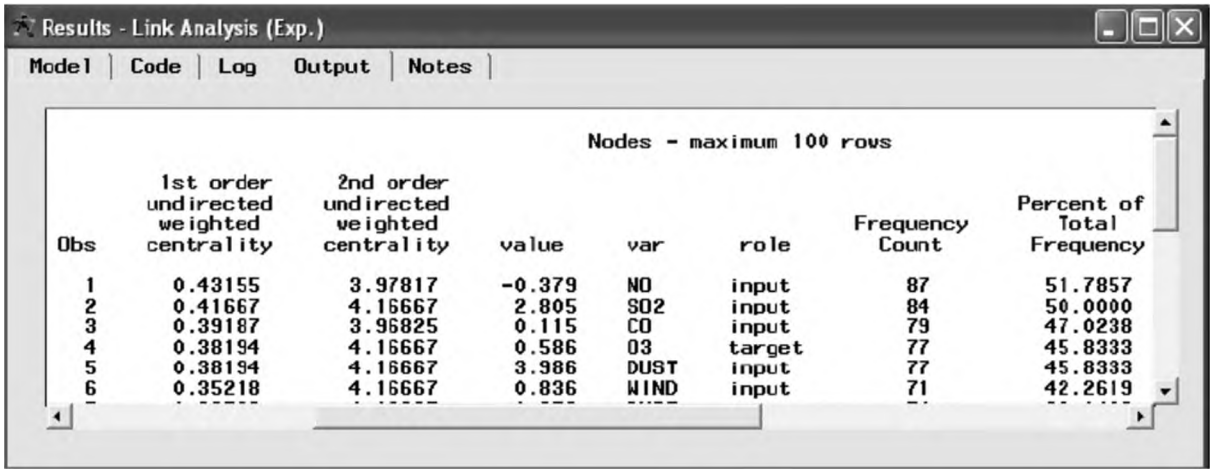


Figura 9-28

Visor de resultados

La pestaña *Output* de la pantalla *Link Analysis* →*Results* (Figura 9-29) arroja los resultados en el visor. En esta ocasión hemos utilizado el archivo AIR relacionando el ozono troposférico (*output*) con el resto de inputs: NO, SO2, CO, DUST y WIND y todas sus posibles interacciones. Los resultados de la ventana *Output* no son guardados en ningún otro archivo.

©Alfaomega-RC Libros



Nodes - maximum 100 rows

Obs	1st order undirected weighted centrality	2nd order undirected weighted centrality	value	var	role	Frequency Count	Percent of Total Frequency
1	0.43155	3.97817	-0.379	NO	input	87	51.7857
2	0.41667	4.16667	2.805	SO2	input	84	50.0000
3	0.39187	3.96825	0.115	CO	input	79	47.0238
4	0.38194	4.16667	0.586	O3	target	77	45.8333
5	0.38194	4.16667	3.986	DUST	input	77	45.8333
6	0.35218	4.16667	0.836	WIND	input	71	42.2619

Figura 9-29

Una vez vistos los resultados principales podemos acceder a los resultados detallados por medio del icono . Existen varias opciones que se detallan a continuación.

Opción Link Graph

Muestra los resultados del análisis de relaciones. Existen distintas formas de presentar el gráfico según el tipo de datos y de variables. Para acceder a ellas en la barra de herramientas seleccionaremos el icono *layout* (Figura 9-30).

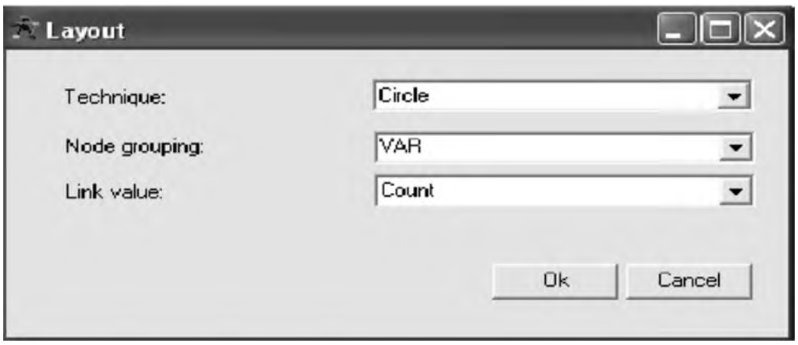


Figura 9-30

Los campos con las opciones son las siguientes:

1) Technique

- ***Circle***: Es la opción por defecto y muestra el gráfico en forma de círculo (Figura 9-31).

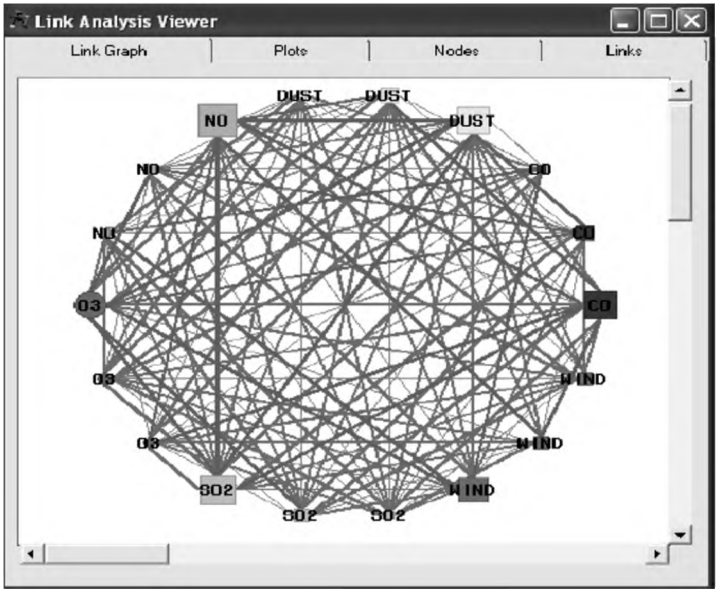


Figura 9-31

Para poner etiquetas a los nodos, sobre el fondo del gráfico activaremos el menú contextual y seleccionaremos *Nodes → Label Text... → Var* (u otra de las opciones). El gráfico representa los niveles de las variables utilizadas (las variables continuas han sido categorizadas en 3 niveles). El gráfico al constar de 6 variables continuas presenta por tanto 18 nodos. Los rasgos del gráfico, comunes a todas las opciones, son los siguientes:

- ***Symbol shape (forma de los nodos)****:* los objetivos se representan en círculos mientras que los inputs se representan por cuadrados.

- ***Symbol color (color de los nodos)****:* cada nodo perteneciente a la misma variable tiene el mismo color.

- ***Symbol size (tamaño del nodo)****:* el tamaño del nodo indica el número de elementos que contiene.

- ***Link width (anchura de la conexión)****:* la anchura de una conexión indica el número de conexiones entre un nodo. Líneas más anchas indica que el nodo en cuestión tiene más conexiones.

©Alfaomega-RC Libros

- ***Link color (color de la conexión)***: los colores son tres. Las conexiones rojas presentan los valores más altos, las azules intermedios y las verdes los más bajos.

- ***Grid***: muestra los resultados en forma de parrilla (Figura 9-32). Es adecuada cuando existe una variable de grupo o cuando los datos han sido agrupados.

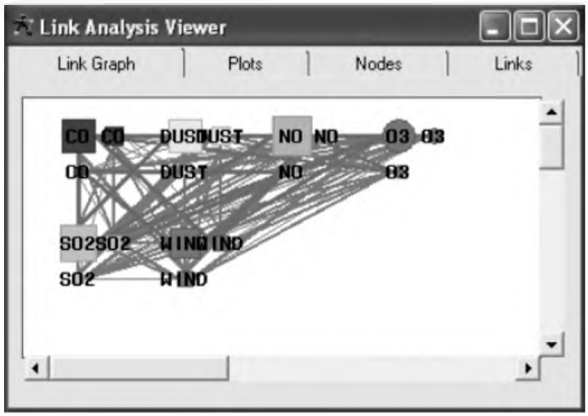


Figura 9-32

- ***MDS***: utiliza los resultados del comando PROC MDS (escalado multidimensional) para escalar los datos en dos dimensiones. Una vez seleccionado MDS se abrirá una ventana con las opciones de la Figura 9-33.

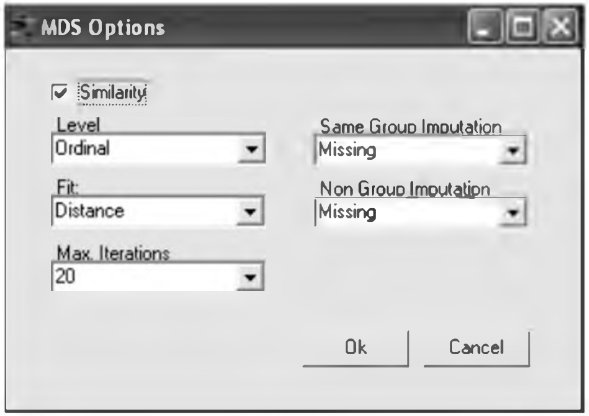


Figura 9-33

- ***Similarity:*** los datos son tratados como medidas de similaridad. Si desactivamos esta opción los trataremos como disimilaridades.

- ***Fit:*** determina las transformaciones a realizar en ambos lados del modelo antes de añadir el término de error. *Distance* ajusta los datos a las distancias, *squared* ajusta los datos al cuadrado a las distancias al cuadrado y *Log* ajusta el logaritmo de los datos al logaritmo de las distancias.

- ***Level:*** especifica el tipo de medida de los datos. Las opciones son Ordinal, Absolute, Ratio, Interval y Log.

- ***Max. Iterations:*** número máximo de iteraciones para encontrar la solución.

- ***Same Group Imputation:*** tratamiento de valores perdidos para observaciones en el mismo grupo.

- ***Non Group Imputation:*** tratamiento de valores perdidos para observaciones no contenidas en el mismo grupo.

Al pulsar *OK* en la Figura 9-33, se obtiene el resultado (Figura 9-34).

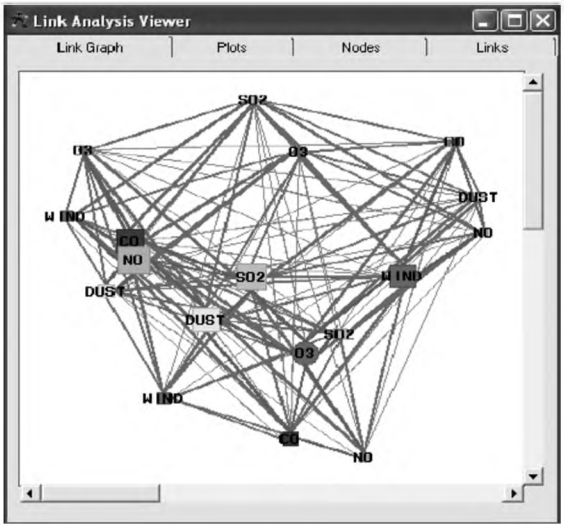


Figura 9-34

©Alfaomega-RC Libros

- ***Parallel Axis***: muestra los nodos en filas o columnas y solo las conexiones adyacentes entre las filas y las columnas (Figura 9-35). Los nodos quedan ordenados de arriba abajo por su importancia al igual que las conexiones. Se permite especificar la orientación del gráfico y el orden de las variables.

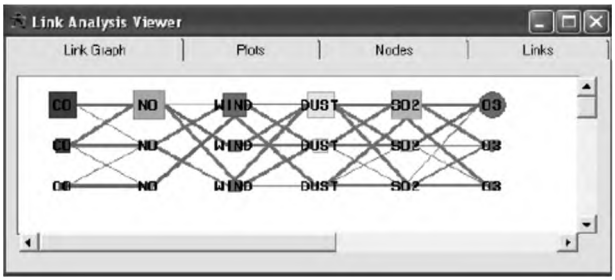


Figura 9-35

- ***Swap:*** permite modificar cualquiera de los gráficos anteriores, intercambiando los nodos de forma que estén más cercanos los nodos con las conexiones más fuertes.

- ***Tree:*** muestra el gráfico con una estructura de árbol usando el comando PROC NETDRAW (Figura 9-36). Podemos especificar la orientación en la ventana *Options* que se abre antes de que el gráfico sea trazado. Las opciones son *Top – Down* (de arriba abajo), *Left – Right* (De izquierda a derecha) y *Polar* (circular).

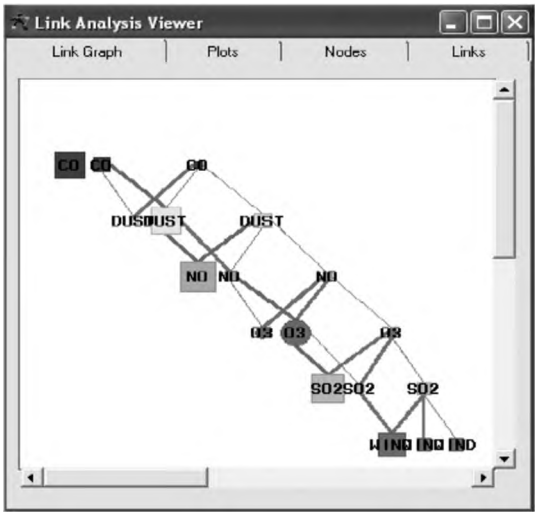


Figura 9-36

2) *Node Grouping*

- ***Value***: los nodos se agrupan en función de su valor.
- ***Var***: los nodos se agrupan por el nombre de su variable. Opción por defecto.
- ***Role:*** los nodos se agrupan a partir de su papel en el modelo.
- ***ID:*** los nodos se agrupan a partir de la etiqueta creada por el análisis para cada nodo.
- ***Text:*** los nodos se agrupan por el texto VAR = Valor.

3) Link Value

- ***Count:*** número de veces que aparece la conexión.
- ***Percent:*** count dividido por el número total de observaciones.
- ***Count1:*** count para ID1.
- ***Count2:*** count para ID2.
- ***Expect:*** el número esperado de conexiones.
- ***Dev:*** la desviación desde el número desperado de conexiones expresado como *Expext* menos *Count*.
- ***Chi2:*** el valor chi-cuadrado de la conexión.

Las opciones gráficas pueden ser variadas a través del menú contextual. Podemos borrar o añadir nodos y conexiones basándonos en distintos criterios, ver información de un nodo o una conexión o variar los colores, formas y tamaños asignados por defecto.

El icono cluster permite agrupar los nodos de forma interactiva o usando técnicas no paramétricas a partir del comando PROC MODECLUS. Las opciones no paramétricas son *kernel density* (especificación esférica con radios fijos) y *nearest neighbor* (especificando un número de vecinos para formar los clusters).

Opción Plots

Permite crear gráficos de barras basados en cualquiera de las variables contenidas en los nodos o en los datos de conexiones (Figura 9-37). Las opciones disponibles permiten modificar el tipo de gráfico.

 ©Alfaomega-RC Libros

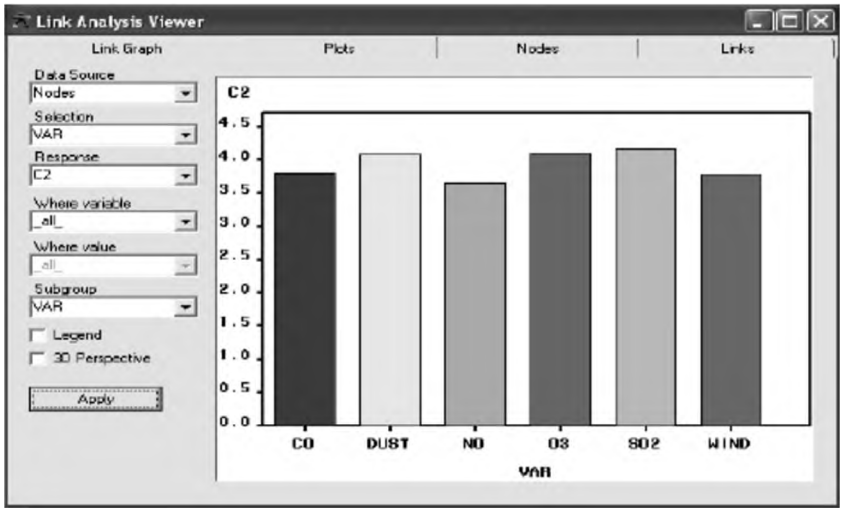


Figura 9-37

Opción Nodes

Muestra las características de los nodos ordenados por su nombre (Figura 9-38).

Link Analysis Viewer

Link Graph | Plots | Nodes | Links

	C1	C2	VALUE	VAR	ROLE	COUNT	PERCENT	ID	TEXT
N1	0.3918650794	3.9682539683	0.115	CO	input	79	47.023809524	N1	CO=0.115
N2	0.2529761905	4.1666666667	1.396	CO	input	51	30.357142857	N2	CO=1.396
N3	0.1884920635	3.253968254	2.677	CO	input	38	22.619047619	N3	CO=2.677
N5	0.3819444444	4.1666666667	3.986	DUST	input	77	45.833333333	N5	DUST=3.986
N4	0.2678571429	4.1666666667	1.358	DUST	input	54	32.142857143	N4	DUST=1.358

Figura 9-38

Las variables que se muestran son las siguientes:

- ***C1***: medida de centralidad de primer orden.
- ***C2***: medida de centralidad de segundo orden.
- ***Value***: valor de la variable de clase o punto medio de las categorías en las que fue dividida la variable continua.
- ***Var***: variable que constituye el nodo.
- ***Role:*** el papel de la variable en el análisis.
- ***Count***: número de observaciones representadas por ese nivel de la variable.
- ***Percent:*** count dividido por el total de observaciones.
- ***ID***: etiqueta identificativa del nodo.
- ***Text***: variable = Valor
- ***X***: la coordenada en el eje de abscisas del gráfico de conexiones.
- ***Y***: la coordenada en el eje de ordenadas del gráfico de conexiones.

Opción Links

Muestra las características de las conexiones ordenadas por su nombre (Figura 9-39).

Link Analysis Viewer

Link Graph | Plots | Nodes | Links

	COUNT	PERCENT	COUNT1	ID1	COUNT2	ID2	LINKID	EXPECT	DEV	CHI2	LA
L30	1	0.5952380952	51	N2	43	N12	L30	13.053571429	-12.05357143	11.130178816	1
L6	1	0.5952380952	38	N3	54	N4	L6	12.214285714	-11.21428571	10.296157059	1
L10	1	0.5952380952	79	N1	37	N6	L10	17.398809524	-16.39880952	15.456284721	1
L28	1	0.5952380952	37	N6	87	N7	L28	19.160714286	-18.16071429	17.212904407	1
L21	1	0.5952380952	54	N4	40	N9	L21	12.857142857	-11.85714286	10.934920635	1

Figura 9-39

Las variables que se muestran son las siguientes:

- ***Count***: número de veces que aparece la conexión.
- ***Percent***: count dividido por el número de observaciones en porcentaje.
- ***Count1***: count para ID1.
- ***ID1***: el primer nodo de la conexión.
- ***Count2***: count para ID2.
- ***ID2***: el segundo nodo de la conexión.
- ***Linkid***: la etiqueta de la conexión.
- ***Expect***: número de veces esperado de aparición de la conexión.
- ***Dev***: Expect menos Count.
- ***Chi2***: valor chi-cuadrado de la conexión.
- ***Layout_Mask***: determina si se aplica la máscara a la conexión. El valor 1 indica que se incluye y el valor 0 indica su exclusión.

©Alfaomega-RC Libros

CAPÍTULO 10

PRONÓSTICOS Y TENDENCIAS CON HERRAMIENTAS DE IBM

EL ALGORITMO SERIE TEMPORAL DE IBM SPSS MODELER

Pronosticar consiste en predecir los valores de una o varias series a lo largo del tiempo. Por ejemplo, puede que desee predecir la demanda esperada de una línea de productos o servicios con la finalidad de poder asignar recursos para su fabricación o distribución. Como para implementar las decisiones de planificación es necesario cierto tiempo, las predicciones son una herramienta esencial en muchos procesos de planificación.

Los métodos de modelado de series temporales suponen que la historia se repite, si no exactamente, de una manera lo suficientemente parecida como para que estudiando el pasado sea posible tomar decisiones mejores en el futuro. Para predecir las ventas del año que viene, por ejemplo, es probable que empiece examinando las ventas de este año y después las de años anteriores para averiguar las tendencias o los patrones, si los hay, que se han desarrollado en los últimos años. No obstante, los patrones pueden ser difíciles de calcular. Si las ventas aumentan

durante varias semanas seguidas, por ejemplo, ¿forma esto parte de un ciclo estacional o se trata del principio de una tendencia a largo plazo?

Con las técnicas de modelado estadístico, puede analizar los patrones de los datos del pasado y proyectar dichos patrones para determinar el rango en el que probablemente se incluirán los valores futuros de la serie. Como resultado, se obtienen predicciones más precisas en las que podrá basar sus decisiones.

Una serie temporal es una colección ordenada de medidas tomadas en intervalos regulares; por ejemplo, los precios diarios de las acciones o los datos de ventas semanales. Las medidas pueden estar relacionadas con cualquier cosa que le interese, y cada serie se suele clasificar en una de las siguientes categorías:

• *Dependiente*. Serie que se desea pronosticar.

• *Predictora*. Serie que puede ayudar a explicar el objetivo, por ejemplo, el presupuesto de publicidad para predecir las ventas. Las series predictoras solo se pueden usar con modelos ARIMA.

• *Evento*. Serie predictora especial que se utiliza para tener en cuenta incidentes recurrentes predecibles como, por ejemplo, las promociones de ventas.

• *Intervención*. Serie predictora especial que se utiliza para tener en cuenta incidentes puntuales del pasado como, por ejemplo, apagones o huelgas.

Los intervalos pueden representar cualquier unidad de tiempo, pero debe utilizarse un mismo intervalo para todas las medidas. Además, si algún intervalo no tiene ninguna medida, debe definirse en el valor perdido. De esta forma, el número de intervalos con medidas (incluidos los que tienen valores perdidos) define la duración del período histórico de los datos.

Componentes de una serie temporal

Estudiar el comportamiento pasado de una serie ayudará a identificar los patrones y realizar mejores pronósticos. Cuando se representan, muchas series temporales muestran una o varias de estas características:

• *Tendencias*

• *Ciclos estacionales y no estacionales*

©Alfaomega-RC Libros

- *Pulsos y pasos*

- *Valores atípicos*

Tendencias

Una tendencia es un cambio gradual ascendente o descendente en el nivel de la serie o la trayectoria que siguen los valores de la serie de aumentar o disminuir a lo largo del tiempo.

Las tendencias pueden ser locales o globales, pero una misma serie puede mostrar ambas. Históricamente, los gráficos de series del índice del mercado de valores muestran una tendencia global ascendente. Han aparecido tendencias descendentes locales en épocas de recesión y tendencias ascendentes locales en épocas de prosperidad.

Las tendencias también pueden ser lineales o no lineales. Las tendencias lineales son incrementos aditivos positivos o negativos en el nivel de la serie, comparables al efecto del interés simple sobre el principal. Las tendencias no lineales suelen ser multiplicativas, con incrementos proporcionales a los valores de series anteriores.

Las tendencias lineales globales se ajustan y pronostican correctamente mediante modelos ARIMA y de suavizado exponencial. Al generar modelos ARIMA, suelen diferenciarse las series que muestran tendencias para eliminar el efecto de estas.

Ciclos estacionales

Un ciclo estacional es un patrón repetitivo y predecible de los valores de las series. Los ciclos estacionales están ligados al intervalo de la serie. Por ejemplo, los datos mensuales suelen mostrar un comportamiento cíclico a lo largo de trimestres y años. Una serie mensual puede mostrar un ciclo trimestral significativo con un mínimo en el primer trimestre o un ciclo anual con un pico en cada mes de diciembre. Se dice que las series con un ciclo estacional muestran estacionalidad.

Los patrones estacionales resultan útiles para obtener buenos ajustes y predicciones. Hay modelos ARIMA y de suavizado exponencial que capturan la estacionalidad.

Ciclos no estacionales

Un ciclo no estacional es un patrón repetitivo y posiblemente impredecible de los valores de las series. Algunas series, como la tasa de desempleo, muestran un claro comportamiento cíclico; no obstante, la periodicidad del ciclo varía a lo largo del tiempo, por lo que resulta difícil predecir cuándo se van a producir máximos o mínimos. Otras series pueden tener ciclos predecibles, pero no se ajustan exactamente al calendario gregoriano o tienen ciclos que se prolongan más de un año. Por ejemplo, las mareas siguen el calendario lunar, los viajes y el comercio internacionales relacionados con los Juegos Olímpicos aumentan cada cuatro años, y hay muchas festividades religiosas cuyas fechas gregorianas cambian de un año a otro.

Los patrones cíclicos no estacionales son difíciles de modelar y suelen aumentar la incertidumbre de los pronósticos. El mercado de valores, por ejemplo, proporciona numerosos ejemplos de series que han desafiado el trabajo de los pronosticadores. No obstante, los patrones no estacionales se deben tener en cuenta cuando existen. En muchos casos, aun así es posible identificar un modelo que se ajuste a los datos históricos razonablemente bien, lo que le ofrece una oportunidad excelente para minimizar la incertidumbre de los pronósticos.

Pulsos y pasos

Muchas series experimentan cambios bruscos de nivel. Normalmente son de dos tipos:

- Un cambio repentino y temporal, o pulso, en el nivel de la serie.

- Un cambio repentino y permanente, o paso, en el nivel de la serie.

Series con pulsos

Cuando se observan pasos o pulsos, es importante encontrar una explicación convincente. Los modelos de series temporales están diseñados para explicar cambios graduales y no repentinos. Por tanto, suelen subestimar los pulsos y pueden quedar inutilizados por los pasos, lo que da como resultado modelos poco ajustados y predicciones imprecisas. (Es posible que algunos casos de estacionalidad parezcan presentar cambios repentinos de nivel, pero que el nivel sea constante de un período estacional a otro).

 ©Alfaomega-RC Libros

Si puede explicarse una alteración, se puede modelar mediante una intervención o un evento. Por ejemplo, en agosto de 1973, la Organización de Países Exportadores de Petróleo (OPEP) impuso un embargo sobre el petróleo que cambió drásticamente la tasa de inflación, aunque recuperó sus niveles normales en los meses siguientes. Si especifica una intervención por puntos para el mes del embargo, puede mejorar el ajuste del modelo, lo que mejorará las predicciones indirectamente. Por ejemplo, puede que un comercio minorista descubra que sus ventas se incrementaron mucho más de lo normal un día que todos los artículos se rebajaron un 50%. Si se especifica una promoción de rebajas del 50% como evento recurrente, puede mejorar el ajuste del modelo y estimar la repercusión que tendría esa misma promoción en el futuro.

Valores atípicos

Los desplazamientos en el nivel de una serie temporal que no se pueden explicar se denominan valores atípicos. Estas observaciones no coinciden con el resto de las series y pueden influir considerablemente en el análisis y, por lo tanto, afectar a la capacidad de pronóstico del modelo de serie temporal. Existen diferentes tipos de valores atípicos:

• *Valor atípico aditivo*. Un valor atípico aditivo aparece como un valor inesperadamente alto o bajo que se produce para una única observación. Las siguientes observaciones no se ven afectadas por un valor atípico aditivo. Los siguientes valores atípicos aditivos se denominan normalmente parches de valores atípicos aditivos.

• *Valor atípico innovador*. Un valor atípico innovador se caracteriza por un impacto inicial con efectos que se extienden sobre las siguientes observaciones. La influencia de los valores atípicos puede aumentar mientras avanza el tiempo.

• *Valor atípico de cambio de nivel*. En el cambio de nivel, todas las observaciones que aparecen después del valor atípico se desplazan a un nuevo nivel. A diferencia de los valores atípicos aditivos, un valor atípico de cambio de nivel afecta a diversas observaciones y tiene un efecto permanente.

• *Valor atípico de cambio transitorio*. Los valores atípicos de cambio transitorio son similares a los valores atípicos de cambio de nivel, pero su efecto se reduce exponencialmente en las siguientes observaciones. Finalmente, las series vuelven a su nivel normal.

• *Valor atípico aditivo estacional*. Un valor atípico aditivo estacional aparece como un valor inesperadamente alto o bajo que se produce repetidamente en intervalos regulares.

• *Valor atípico de tendencia local*. Un valor atípico de tendencia local produce un cambio general en la serie causado por un patrón en los valores atípicos después de la aparición del valor atípico inicial.

La detección de valores atípicos en una serie temporal implica determinar la ubicación, tipo y magnitud de todos los valores atípicos presentes. Tsay (1998) propuso un procedimiento iterativo para detectar el cambio del nivel de la media con el fin de identificar los valores atípicos deterministas. Este proceso implica la comparación de un modelo de serie temporal que supone que no hay presentes valores atípicos con otro modelo que incorpore valores atípicos. Las diferencias entre modelos permiten calcular el efecto de tratar cualquier punto como un valor atípico.

Funciones de autocorrelación y autocorrelación parcial

La autocorrelación y la autocorrelación parcial son medidas de asociación entre valores de series actuales y pasadas e indican cuáles son los valores de series pasadas más útiles para predecir valores futuros. Con estos datos podrá determinar el orden de los procesos en un modelo ARIMA. Más concretamente:

• Función de autocorrelación (FAS). En el retardo k, es la autocorrelación entre los valores de las series que se encuentran a k intervalos de distancia.

• Función de autocorrelación parcial (FAP). En el retardo k, es la autocorrelación entre los valores de las series que se encuentran a k intervalos de distancia, teniendo en cuenta los valores de los intervalos intermedios.

Gráfico de FAS de una serie

El eje X del gráfico de FAS indica el retardo en el que se calcula la autocorrelación; el eje Y indica el valor de la correlación (entre −1 y 1). Por ejemplo, un trazo de unión en el retardo 1 de un gráfico de FAS indica que existe una fuerte correlación entre el valor de cada serie y el valor anterior, un trazo de unión en el retardo 2 indica que existe una fuerte correlación entre el valor de cada serie y el valor que aparece dos puntos anteriores, etc.

 ©Alfaomega-RC Libros

• Una correlación positiva indica que los valores grandes actuales se corresponden con valores grandes en el retardo especificado; una correlación negativa indica que los valores grandes actuales se corresponden con valores pequeños en el retardo especificado.

• El valor absoluto de una correlación es una medida de la fuerza de la asociación, con valores absolutos mayores que indican relaciones más fuertes.

Transformaciones de series

Las transformaciones suelen ser útiles para estabilizar una serie antes de estimar modelos. Esto es especialmente importante para modelos ARIMA, que necesitan que las series sean estacionarias antes de estimar los modelos. Una serie es estacionaria si el nivel global (media) y la desvianza media del nivel (varianza) son constantes a lo largo de la serie.

Aunque la mayoría de las series interesantes no son estacionarias, ARIMA es eficaz siempre y cuando la serie se pueda convertir en estacionaria mediante la aplicación de transformaciones tales como el logaritmo natural, la diferenciación o la diferenciación estacional.

Transformaciones de estabilización de la varianza. Las series en las que la varianza cambia a lo largo del tiempo con frecuencia se pueden estabilizar con una transformación logarítmica natural o de raíz cuadrada. También reciben el nombre de transformaciones funcionales.

• *Log natural*. El logaritmo natural se aplica a los valores de las series.

• *Raíz cuadrada*. La función de raíz cuadrada se aplica a los valores de las series.

No se pueden usar las transformaciones logarítmica natural o de raíz cuadrada para series con valores negativos.

Transformaciones de estabilización del nivel. Un suave descenso de los valores de la FAS indica que todos los valores de la serie están estrechamente correlacionados con el valor anterior. Si analiza el cambio de los valores de la serie, obtendrá un nivel estable.

• *Diferenciación simple*. Se calculan las diferencias existentes entre cada valor y el anterior de la serie, a excepción del valor más antiguo de la serie. Por tanto, la serie diferenciada tendrá un valor menos que la serie original.

• *Diferenciación estacional*. Es idéntica a la diferenciación simple, excepto en que se calculan las diferencias existentes entre cada valor y el valor estacional anterior.

Si se usa la diferenciación simple o estacional de forma simultánea con la transformación logarítmica o de raíz cuadrada, siempre se aplicará primero la transformación de estabilización de la varianza. Si se usan la diferenciación simple y la estacional, los valores de la serie resultante son iguales independientemente de si se aplica primero una diferenciación u otra.

Serie predictora

La serie predictora contiene datos relacionados que pueden ayudar a explicar el comportamiento de la serie que se va a pronosticar. Por ejemplo, un minorista de venta por catálogo o por Internet podría predecir el número de ventas en función del número de catálogos enviados, el número de líneas telefónicas abiertas o el número de entradas a la página Web de su empresa.

Cualquier serie puede utilizarse como un predictor siempre que se extienda en el tiempo que desea pronosticar y tenga los datos completos, sin valores perdidos.

Tenga cuidado al añadir predictores a un modelo, ya que añadir un gran número de predictores aumentará el tiempo necesario para calcular los modelos. Aunque añadir predictores puede mejorar la capacidad del modelo para ajustarse a los datos históricos, no significa necesariamente que el modelo vaya a realizar un mejor pronóstico, por lo que una mayor complejidad puede no valer la pena. Lo ideal sería identificar el modelo más simple que mejores pronósticos realice.

Como norma general, se recomienda que el número de predictores sea inferior al tamaño muestral dividido entre 15 (como mucho, un predictor por 15 casos).

Los predictores con datos incompletos o perdidos no pueden utilizarse para la predicción. Esto es aplicable tanto a los datos históricos como a los valores futuros. En algunos casos, puede evitar esta limitación mediante la configuración de la amplitud de estimación del modelo para excluir los datos más antiguos a la hora de calcular los modelos.

©Alfaomega-RC Libros

Nodo Modelos de series temporales

El nodo Serie temporal estima modelos de suavizado exponencial, modelos autorregresivos integrados de media móvil (ARIMA) univariados y modelos ARIMA (o de función de transferencia) multivariados para series temporales y genera predicciones a partir de los datos de series temporales.

El suavizado exponencial es un método de predicción que utiliza los valores ponderados de las observaciones anteriores de la serie para pronosticar los valores futuros. Como tal, el suavizado exponencial no se basa en una comprensión teórica de los datos. Pronostica un punto cada vez, corrigiendo las predicciones a medida que entran nuevos datos. Esta técnica es útil para pronosticar las series que muestran una tendencia, estacionalidad o ambas. Puede elegir entre una amplia variedad de modelos de suavizado exponencial que difieren en el tratamiento de la tendencia y la estacionalidad.

Los modelos ARIMA proporcionan métodos más sofisticados para crear modelos de los componentes de tendencia y estacionales que los modelos de suavizado exponencial y, en concreto, ofrecen la ventaja adicional de incluir variables independientes (predictoras) en el modelo. Esto implica la especificación explícita de órdenes autorregresivos y de media móvil además del grado de diferenciación. Puede incluir variables del predictor y definir funciones de transferencia para algunas o todas ellas, así como especificar la detección automática de valores atípicos o especificar un conjunto explícito de valores atípicos.

Nota: en términos prácticos, los modelos ARIMA son especialmente útiles si desea incluir predictores que puedan ayudar a explicar el comportamiento de la serie que se está pronosticando, como el número de catálogos enviados por correo o el número de visitas de la página Web de una empresa. Los modelos de suavizado exponencial describen el comportamiento de la serie temporal sin tratar de comprender el motivo de su comportamiento. Por ejemplo, una serie que históricamente ha mostrado picos cada 12 meses, es probable que lo siga haciendo aunque se desconozca el motivo.

Existe también un modelizador experto que identifica y estima automáticamente el modelo ARIMA o de suavizado exponencial que mejor se ajusta para una o más variables objetivo, lo que elimina la necesidad de identificar un modelo adecuado mediante ensayo y error. En todos los casos, el modelizador experto elige el mejor modelo para cada variable objetivo especificada. En caso de duda, utilice el modelizador experto.

Si se especifican variables predictoras, el modelizador experto selecciona, para su inclusión en los modelos ARIMA, las variables que tienen una relación estadísticamente significativa con la serie dependiente. Las variables del modelo se transforman cuando es necesario mediante una diferenciación y/o una raíz cuadrada o una transformación logarítmica natural. Por defecto, el modelizador experto tiene en cuenta todos los modelos de suavizado exponencial y todos los modelos ARIMA y elige el mejor modelo para cada campo objetivo. Sin embargo, puede limitar el modelizador experto para que solo elija el mejor modelo de suavizado exponencial o para que solo elija el mejor modelo ARIMA. Además, puede especificar la detección automática de valores atípicos.

El nodo *Serie temporal* se distingue de otros nodos de IBM® SPSS® Modeler en que no se puede insertar simplemente en una ruta y ejecutar dicha ruta. El nodo *Serie temporal* debe ir siempre precedido por un nodo *Intervalos de tiempo* que especifique información como el intervalo de tiempo que se va a utilizar (años, trimestres, meses, etc.), los datos que se van a utilizar para la estimación y lo que se va a prolongar un pronóstico en el futuro, si se utiliza.

Los datos de serie temporal deben estar espaciados de manera uniforme. Los métodos para modelar datos de series temporales precisan de un intervalo uniforme entre cada medida. Los valores perdidos deben indicarse mediante filas vacías. Si sus datos aún no cumplen este requisito, el nodo Intervalos de tiempo puede transformar los valores según sea necesario.

Deben tenerse en cuenta estas otras cuestiones en relación con el nodo Serie temporal:

- Los campos deben ser numéricos.
- Los campos de fecha no se pueden utilizar como entradas.
- Las particiones se ignorarán.

Como ejemplo consideramos un modelo que trata de predecir el valor futuro durante los próximos meses de los ingresos por impuestos en IRPF, IVA y SOCIEDADES en el sistema impositivo español, utilizando como variable predictora los ingresos totales por impuestos (variable ITRI).

Una vez importado el fichero que contiene los datos de nombre *Taxes_2011.sav* como un origen de datos IBM Statistics, utilizamos el nodo *Intervalo de tiempo* de la ficha *Operaciones con campos* para fechar las series temporales del fichero origen de datos. En la solapa *Intervalos* (Figura 10-1) definimos el *Intervalo de*

 ©Alfaomega-RC Libros

tiempo como *Meses* ya que la serie es mensual y definimos su inicio en Enero del año 2000. En la solapa *Estimación* (Figura 10-2) fijamos las estimaciones utilizando todo el período de la serie desde el principio hasta el final. En la solapa *Predicción* (Figura 10-3) fijamos 12 meses de predicciones. Hacemos clic en *Aplicar* y *Aceptar*.

A continuación utilizamos el nodo *Tipo* para asignar a la variable ITRI el papel de *Entrada,* ya que va a ser la variable predictora del modelo. También suele asignarse el papel de *Entrada* a la variable de tiempo *Date*, ya que el tiempo es la variable independiente en una serie temporal. El mismo papel se asignará a todas las variables candidatas a ser predictoras, como por ejemplo INNO. A las variables de impuestos que van a ser predichas se les asigna el papel de variables *objetivo* (Figura 10-4). Se hace clic en *Aplicar* y *Aceptar*.

Figura 10-1

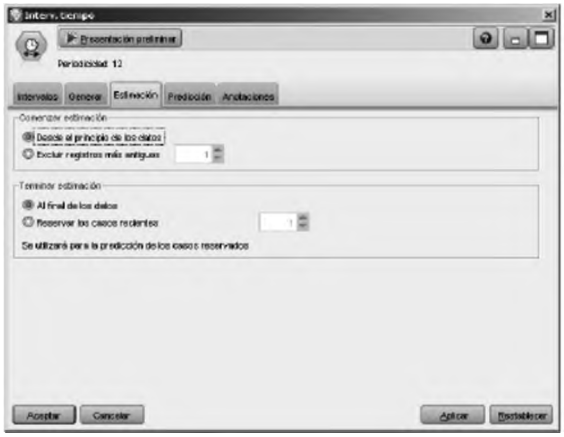


Figura 10-2

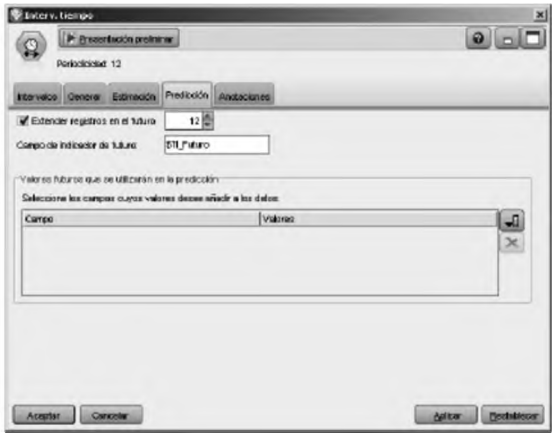


Figura 10-3

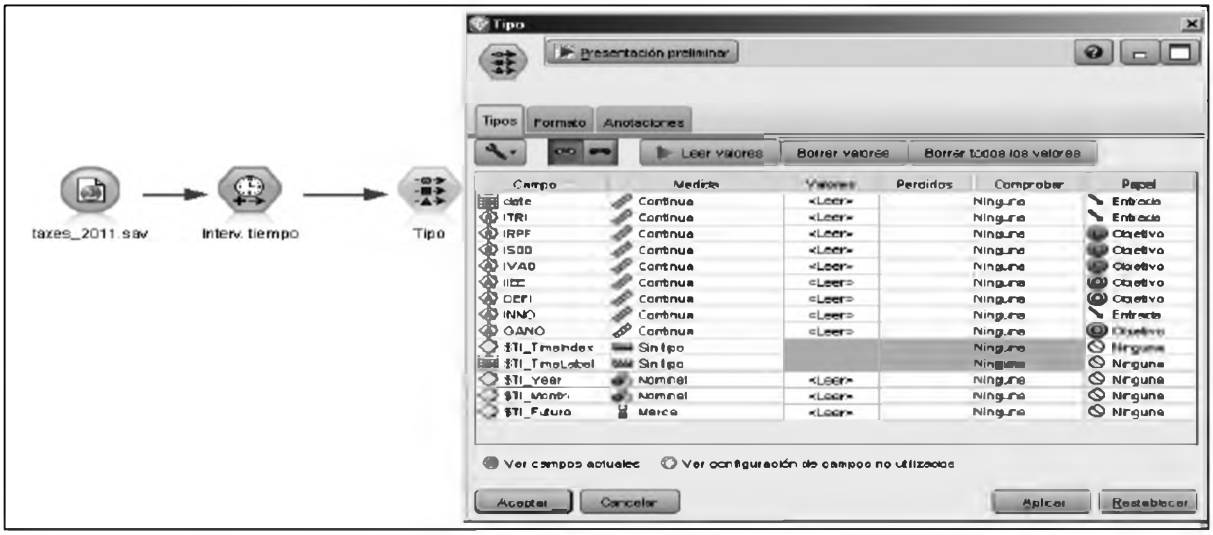


Figura 10-4

A continuación se une al diagrama el nodo *Serie temporal*, se hace doble clic sobre él con el botón secundario del ratón y se elige *Edición* en el menú emergente resultante para definir los parámetros en la Figura 10-5. En la solapa *Campos* elegimos *Utilizar configuración personalizada* para definir las tres series que vamos a predecir como series *Objetivo* y la serie predictora como serie de *Entrada*. En la solapa *Modelo* (Figura 10-6) elegimos *Automático* y el *Modelizador experto* como *Método*, fijamos la *Amplitud del límite de confianza* al 95% y 48 términos para la FAS y la FAP. Al hacer clic en el botón *Criterios* de la Figura 10-6 se obtiene la Figura 10-7 o para fijar criterios del modelizador experto. En la solapa *Modelos* fijamos *Todos los modelos* y en la solapa *Valores atípicos* (Figura 10-8) fijamos toda la tipología posible con detección automática. Al hacer clic en *Aplicar* y *Ejecutar* se obtiene el modelo sobre la ruta (Figura 10-9).

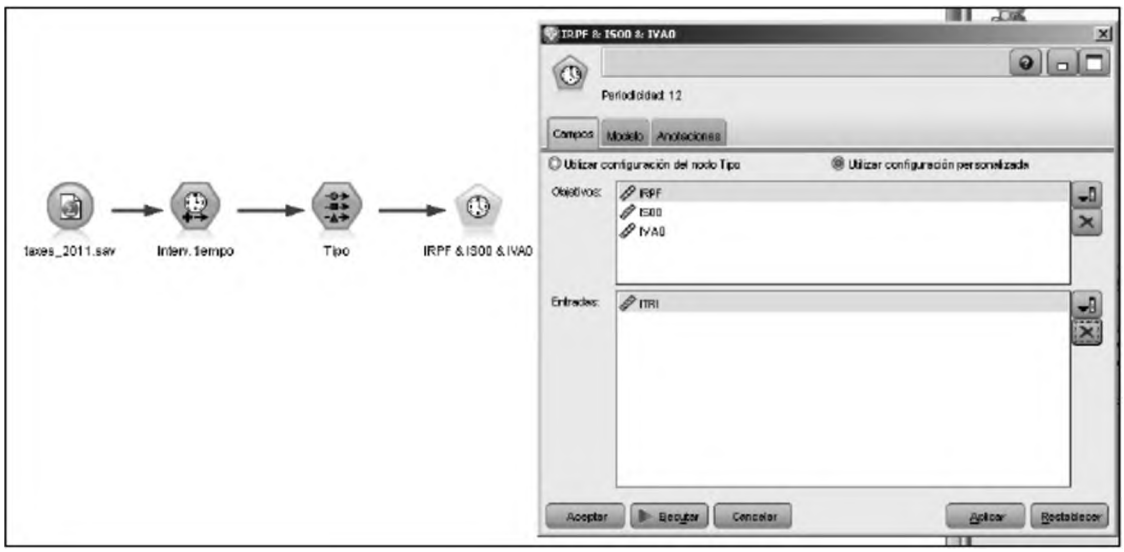


Figura 10-5

©Alfaomega-RC Libros

Figura 10-6

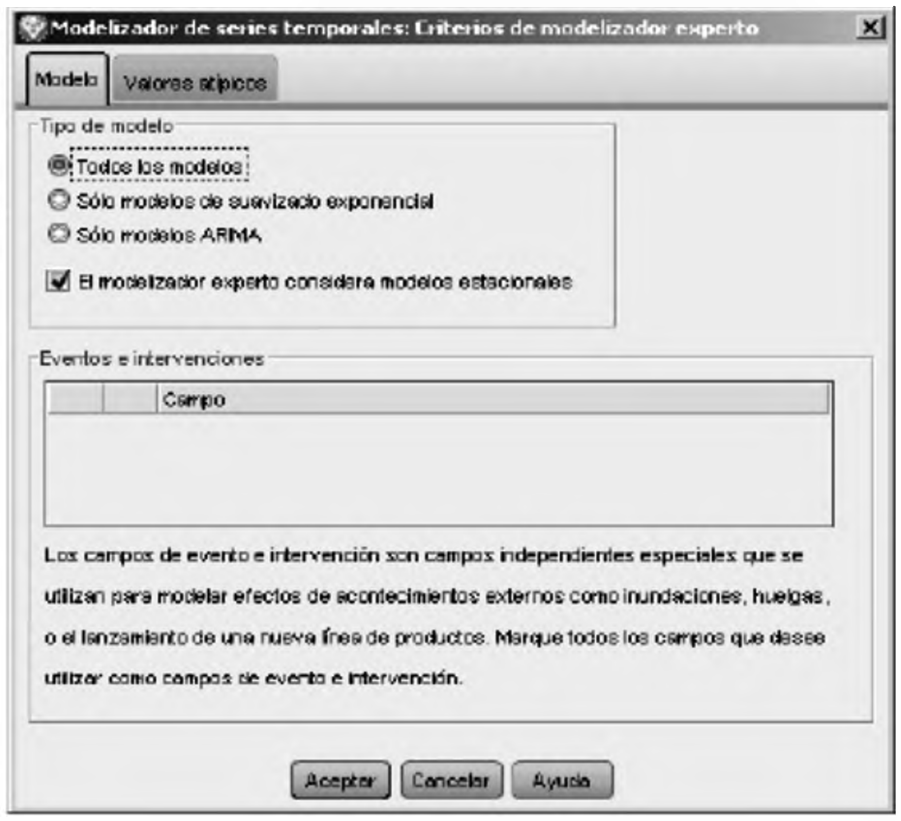


Figura 10-7

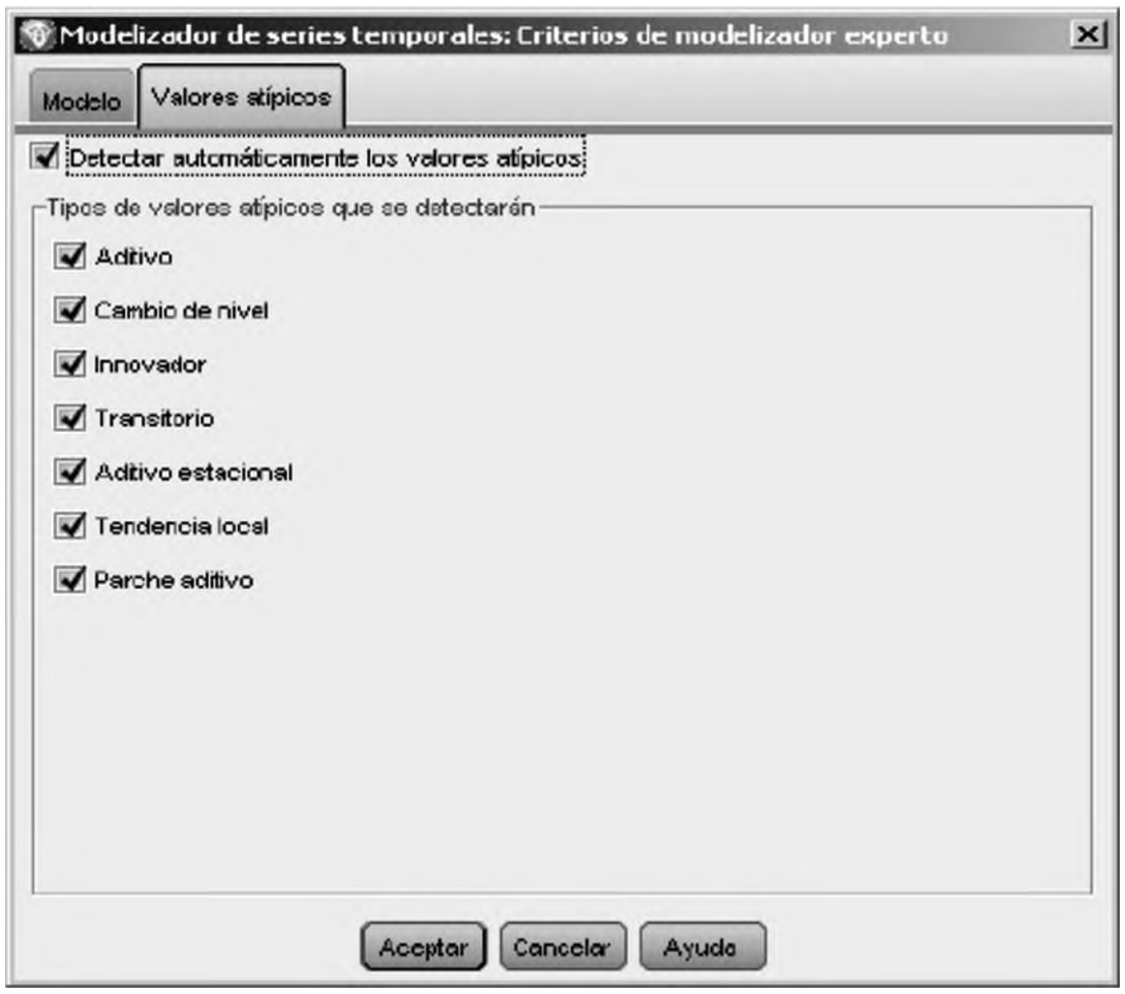


Figura 10-8

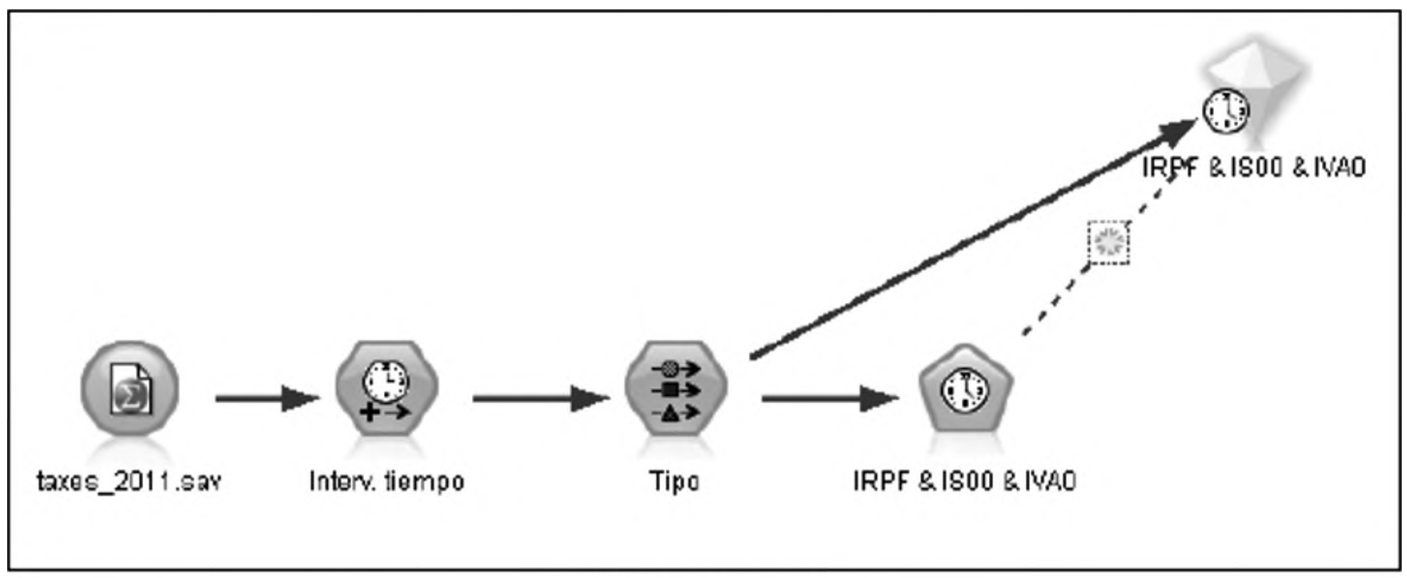


Figura 10-9

En la pestaña *Campos* (Figura 10-5) se especifican los campos que se van a utilizar para generar el modelo. Para generar un modelo, primero se deben especificar los campos que se desea usar como objetivos y como entradas. Como hemos visto, el nodo *Serie temporal* suele utilizar información de campo de un nodo Tipo situado en un punto anterior de la ruta. Si utiliza un nodo Tipo para seleccionar campos de entrada y objetivo, no es necesario cambiar nada en esta pestaña.

Utilizar configuración del nodo Tipo. Esta opción permite indicar al nodo que use la información de campo de un nodo Tipo situado en un punto anterior de la ruta. Este es el método por defecto.

Utilizar configuración personalizada. Esta opción permite indicar al nodo que use la información de campo especificada aquí en lugar de la proporcionada en nodos Tipo situados en cualquier punto anterior de la ruta. Después de seleccionar esta opción, especifique los campos siguientes. Tenga en cuenta que los campos almacenados como fechas no se aceptan como campos objetivo o de entrada.

• *Objetivos*. Seleccione uno o varios campos objetivo. Se trata de una acción similar a establecer un papel de un campo a Objetivo en un nodo Tipo. Los campos objetivos de un modelo de series temporales debe tener un nivel de medición de Continuo. Se crea un modelo distinto para cada campo objetivo. Un campo objetivo tiene en cuenta todos los campos Entrada especificados (excepto dicho campo) como posibles entradas. Por lo tanto, el mismo campo puede incluirse en ambas listas; ese campo se utilizará como posible entrada para todos los modelos, excepto en el que actúa como objetivo.

• *Entradas*. Seleccione los campos de entrada. Se trata de una acción similar a establecer un papel de un campo a Entrada en un nodo Tipo. Los campos de entrada de un modelo de serie temporal deben ser numéricos.

En la pestaña *Modelo* (Figura 10-10) se fijan las opciones del modelo de serie temporal:

Nombre del modelo. Especifica el nombre asignado al modelo generado al ejecutar el nodo.

• *Automático*. Genera el nombre del modelo de forma automática basándose en los nombres de los campos objetivo o de ID, o en el nombre del tipo de modelo en los casos en los que no se especifique ningún campo objetivo (como en los modelos de conglomerado).

 ©Alfaomega-RC Libros

• *Personalizado*. Permite especificar un nombre personalizado para el nugget de modelo.

Continuar con la estimación utilizando modelo(s) existente. Si ya ha generado un modelo de serie temporal, seleccione esta opción para reutilizar la configuración de criterios especificada para ese modelo y generar un nuevo nodo de modelo en la paleta de modelos, en lugar de crear uno a partir de cero. De esta manera podrá ahorrar tiempo si vuelve a estimar y generar un nuevo pronóstico basado en la misma configuración del modelo de antes, pero con datos más recientes. Así, si el modelo original de una serie temporal determinada era, por ejemplo, Tendencia lineal de Holt, se utilizará el mismo tipo de modelo para volver a estimar esos datos y realizar un pronóstico con ellos; el sistema no volverá a intentar buscar el mejor tipo de modelo para los nuevos datos. Si selecciona esta opción, desactivará los controles Método y Criterios.

Método. Puede elegir Modelizador experto, Suavizado exponencial o ARIMA.

• *Modelizador experto*. Seleccione esta opción para utilizar el modelizador experto, que busca automáticamente el modelo que mejor se ajusta a cada serie dependiente.

• *Suavizado exponencial*. Utilice esta opción para especificar un modelo de suavizado exponencial personalizado.

• ARIMA. Utilice esta opción para especificar un modelo ARIMA personalizado.

Información de intervalo de tiempo. Esta sección del cuadro de diálogo contiene información sobre las especificaciones para las estimaciones y las predicciones realizadas en el nodo Intervalos de tiempo. Tenga en cuenta que esta sección no aparecerá si elige la opción *Continuar* con la estimación utilizando modelo(s) existente.

En la primera línea se indica si hay algún registro que se excluya del modelo o se utilice como caso reservado. Consulte el tema Período de estimación si desea obtener más información.

En la segunda línea aparece información sobre todos los períodos de predicción especificados en el nodo Intervalos de tiempo.

Si en la primera línea aparece *No se ha definido ningún intervalo de tiempo*, esto significa que no hay ningún nodo Intervalos de tiempo conectado. Esta situación

provocará un error al intentar ejecutar la ruta, por lo que deberá incluir un nodo Intervalos de tiempo en un punto de la ruta anterior al nodo Serie temporal.

Amplitud de límite de confianza (%). Los intervalos de confianza se calculan para las predicciones del modelo y las autocorrelaciones residuales. Puede especificar cualquier valor positivo inferior a 100. Por defecto, se utiliza un intervalo de confianza del 95%.

Número máximo de retardos en resultados de las FAS y FAP. Puede establecer el número máximo de retardos que se muestran en las tablas y en los gráficos de autocorrelaciones y autocorrelaciones parciales.

Construir únicamente modelo de puntuación. Marque esta casilla para reducir la cantidad de datos que están almacenados en el modelo. Al hacerlo mejorará el rendimiento al construir modelos con cifras muy grandes de series temporales (cientos de miles). Si selecciona esta opción, las pestañas Modelo, Parámetros y Residuos no se muestran en el nugget de modelo Series temporales, pero puede puntuar los datos de la manera habitual.

Seleccione *Criterios* para especificar los *Criterios del modelizador experto de series temporales* (Figura 10-11). En la pestaña *Modelo* tenemos los siguientes campos:

Tipo de modelo. Se encuentran disponibles las siguientes opciones:

• *Todos los modelos*. El modelizador experto tiene en cuenta tanto los modelos ARIMA como los modelos de suavizado exponencial.

• *Solo modelos de suavizado exponencial*. El modelizador experto solo tiene en cuenta los modelos de suavizado exponencial.

• *Solo modelos* ARIMA. El modelizador experto solo tiene en cuenta los modelos ARIMA.

El modelizador experto considera modelos estacionales. Esta opción solo está activada si se ha definido una periodicidad para el conjunto de datos activo. Si esta opción está seleccionada, el modelizador experto tiene en cuenta los modelos tanto estacionales como no estacionales. Si esta opción no está seleccionada, el modelizador experto solo tiene en cuenta los modelos no estacionales.

 ©Alfaomega-RC Libros

Eventos e intervenciones. Permite designar determinados campos de entrada como campos de eventos o intervención. De esta forma, se identifica el campo que contiene datos de series temporales afectados por eventos (situaciones recurrentes predecibles, como promociones de ventas) o intervenciones (incidentes puntuales, como apagones o huelgas). El modelizador experto solo tendrá en cuenta la regresión simple y no las funciones de transferencia arbitrarias para entradas identificadas como campos de evento o intervención.

Los campos de entrada deben tener un nivel de medición de Marca, Nominal u Ordinal y deben ser numéricos (por ejemplo, 1/0, no True/False, para un campo marca), antes de que aparezcan en esta lista.

En la pestaña *Valores atípicos* (Figura 10-12) tenemos los sigueinmtes campos:

Detectar automáticamente los valores atípicos. Por defecto, no se realiza la detección automática de valores atípicos. Seleccione esta opción para realizar una detección automática de valores atípicos y, a continuación, seleccione los tipos de valores atípicos que desee. Se pueden marcar también los tipos de valores atípicos que se deen detectar.

Sin en el campo *Método* de la solapa *Modelo* de la Figura 10-10 elegimos *Suavizado exponencial*, al hacer clic en *Criterios* se obtiene la pantalla *Modelizador de series temporales* que permite elegir *Tipo de modelo* y posible *Transformación de objetivo* (Figura 10-14).

Figura 10-10

Tipo de modelo. Los modelos de suavizado exponencial se clasifican como estacionales o no estacionales. Los modelos estacionales solo están disponibles si se ha definido una periodicidad estacional mediante el nodo Intervalos de tiempo. Las periodicidades estacionales son las siguientes: períodos cíclicos, años, trimestres, meses, días por semana, horas por día, minutos por día y segundos por día. Consulte el tema Nodo de intervalos de tiempo si desea obtener más información.

• *Simple*. Este modelo es adecuado para las series sin tendencia ni estacionalidad. Su único parámetro de suavizado relevante es el nivel. El suavizado exponencial simple es muy similar a un ARIMA con cero órdenes de autorregresión, un orden de diferenciación, un orden de media móvil y ninguna constante.

• *Tendencia lineal de Holt*. Este modelo es adecuado para las series con una tendencia lineal y sin estacionalidad. Sus parámetros de suavizado relevantes son el nivel y la tendencia y, en este modelo, no están restringidos por sus valores respectivos. El modelo de Holt es más general que el de Brown, pero puede tardar más en calcular estimaciones para series grandes. El suavizado exponencial de Holt es muy similar a un ARIMA con cero órdenes de autorregresión, dos órdenes de diferenciación y dos órdenes de media móvil.

• *Tendencia lineal de Brown*. Este modelo es adecuado para las series con una tendencia lineal y sin estacionalidad. Sus parámetros de suavizado relevantes son el nivel y la tendencia, pero, en este modelo, se supone que son iguales. Por ello, el modelo de Brown es un caso especial del modelo de Holt. El suavizado exponencial de Brown es muy similar a un ARIMA con cero órdenes de autorregresión, dos órdenes de diferenciación y dos órdenes de media móvil, siendo el coeficiente del segundo orden de la media móvil igual a la mitad del coeficiente del primer orden al cuadrado.

• *Tendencia amortiguada*. Este modelo es adecuado para las series con una tendencia lineal que va desapareciendo y sin estacionalidad. Sus parámetros de suavizado relevantes son el nivel, la tendencia y la tendencia de amortiguación. El suavizado exponencial amortiguado es muy similar a un ARIMA con cero órdenes de autorregresión, un orden de diferenciación y dos órdenes de media móvil.

• *Estacional simple*. Este modelo es adecuado para las series sin una tendencia y un efecto estacional constante a lo largo del tiempo. Sus parámetros de suavizado relevantes son el nivel y la estacionalidad. El suavizado exponencial estacional es muy similar a un ARIMA con cero órdenes de autorregresión, un orden de diferenciación, un orden de diferenciación estacional y los órdenes 1, p y p+1 de media móvil, donde

 ©Alfaomega-RC Libros

p es el número de períodos contenidos en un intervalo estacional. En el caso de los datos mensuales, p = 12.

• *Aditivo de Winters*. Este modelo es adecuado para las series con una tendencia lineal y un efecto estacional constante a lo largo del tiempo. Sus parámetros de suavizado relevantes son el nivel, la tendencia y la estacionalidad. El suavizado exponencial aditivo de Winters es muy similar a un ARIMA con cero órdenes de autorregresión, un orden de diferenciación, un orden de diferenciación estacional y los órdenes p+1 de media móvil, donde p es el número de períodos contenidos en un intervalo estacional. En el caso de los datos mensuales, p = 12.

• *Multiplicativo de Winters*. Este modelo es adecuado para series en las que haya una tendencia lineal y con un efecto estacional que cambie en función de la magnitud de las series. Sus parámetros de suavizado relevantes son el nivel, la tendencia y la estacionalidad. El modelo de suavizado exponencial multiplicativo de Winters no es similar a ningún modelo ARIMA.

Transformación de objetivo. Puede especificar una transformación para que se lleve a cabo en cada variable dependiente antes de su modelado.

• *Ninguno*. No se lleva a cabo ninguna transformación.

• *Raíz cuadrada*. Se realiza una transformación de raíz cuadrada.

• *Log natural*. Se realiza una transformación logarítmica natural.

Criterios ARIMA de series temporales

El nodo Serie temporal le permite generar modelos ARIMA estacionales y no estacionales personalizados (también conocidos como modelos Box-Jenkins) con o sin un conjunto fijo de variables de entrada (predictoras). Puede definir funciones de transferencia para algunas o todas las variables de entrada y especificar la detección automática de valores atípicos o especificar un conjunto explícito de valores atípicos.

Todas las variables de entrada especificadas se incluyen en el modelo de manera explícita, a diferencia de lo que ocurre al utilizar el modelizador experto, donde las variables de entrada solo se incluyen si tienen una relación estadísticamente significativa con la variable objetivo.

Sin en el campo *Método* de la solapa *Modelo* de la Figura 10-10 elegimos *Modelos ARIMA* y hacemos clic en *Criterios*, se obtiene la pantalla *Modelizador de series temporales*: Criterios ARIMA (Figura 10-15) que permite elegir los parámetros de identificación y otras características del modelo ARIMA.

Figura 10-11

La pestaña *Modelo* de la Figura 10-11 le permite especificar la estructura de un modelo ARIMA personalizado. Tenemos los siguientes campos:

Órdenes ARIMA. Escriba valores para los distintos componentes ARIMA del modelo en las casillas correspondientes de la cuadrícula Estructura. Todos los valores deben ser enteros no negativos. Para los componentes autorregresivos y de media móvil, el valor representa el orden máximo. Todos los órdenes inferiores positivos se incluyen en el modelo. Por ejemplo, si especifica 2, el modelo incluye los órdenes 2 y 1. Las casillas de la columna Estacional solo se activan si se ha definido una periodicidad para el conjunto de datos activo.

• Autorregresivo (p). Es el número de órdenes autorregresivos del modelo. Los órdenes autorregresivos especifican los valores previos de la serie utilizados para predecir los valores actuales. Por ejemplo, un orden autorregresivo igual a 2

 ©Alfaomega-RC Libros

especifica que se van a utilizar los valores de la serie correspondientes a dos períodos de tiempo del pasado para predecir el valor actual.

• Diferencia (d). Especifica el orden de diferenciación aplicado a la serie antes de estimar los modelos. La diferenciación es necesaria si hay tendencias (las series con tendencias suelen ser no estacionarias y el modelado de ARIMA asume la estacionariedad) y se utiliza para eliminar su efecto. El orden de diferenciación se corresponde con el grado de la tendencia de la serie (la diferenciación de primer orden representa las tendencias lineales, la diferenciación de segundo orden representa las tendencias cuadráticas, etc.).

• Media móvil (q). Es el número de órdenes de media móvil presentes en el modelo. Los órdenes de media móvil especifican el modo en que se utilizan las desviaciones de la media de la serie para los valores previos con el fin de predecir los valores actuales. Por ejemplo, los órdenes de media móvil de 1 y 2 especifican que las desviaciones del valor medio de la serie de cada uno de los dos últimos períodos de tiempo se tienen en cuenta al predecir los valores actuales de la serie.

Órdenes estacionales. Los componentes estacionales autorregresivos, de media móvil y de diferenciación tienen la misma función que los componentes no estacionales correspondientes. No obstante, en el caso de los órdenes estacionales, los valores de la serie actual se ven afectados por los valores de la serie anterior separados por uno o más períodos estacionales. Por ejemplo, para los datos mensuales (período estacional de 12), un orden estacional de 1 significa que el valor de la serie actual se ve afectado por el valor de la serie 12 períodos antes del actual. Un orden estacional de 1 para los datos mensuales equivale a la especificación de un orden no estacional de 12.

Transformación de objetivo. Puede especificar una transformación para que se lleve a cabo en cada variable objetivo antes de su modelado. Consulte el tema Transformaciones de series si desea obtener más información.

• *Ninguno*. No se lleva a cabo ninguna transformación.

• *Raíz cuadrada*. Se realiza una transformación de raíz cuadrada.

• *Log natural*. Se realiza una transformación logarítmica natural.

Incluir constante en el modelo. La inclusión de una constante es estándar a menos que esté seguro de que el valor de la media global de la serie es 0. Se recomienda la exclusión de la constante si se aplica la diferenciación.

La pestaña *Funciones de transferencia* (Figura 10-12) le permite definir funciones de transferencia para algunos o todos los campos de entrada. Las funciones de transferencia le permiten especificar el modo en que se utilizan los valores anteriores de estos campos para predecir valores futuros de la serie objetivo.

Figura 10-12

Figura 10-13

La pestaña *Funciones de transferencia* se muestra únicamente si los campos de entrada (con el papel definido como Entrada) se especifican, ya sea en el nodo Tipo o en la pestaña *Campos* del nodo Series temporales (seleccione *Utilizar configuración personalizada—Entradas*). En la lista superior aparecen todos los campos de entrada. El resto de la información que aparece en este cuadro de diálogo es específica del campo de entrada seleccionado en la lista.

Órdenes de la función de transferencia. Escriba valores para los distintos componentes de la función de transferencia en las casillas correspondientes de la

 ©Alfaomega-RC Libros

cuadrícula Estructura. Todos los valores deben ser enteros no negativos. Para los componentes de numerador y denominador, el valor representa el orden máximo. Todos los órdenes inferiores positivos se incluyen en el modelo. Además, el orden 0 siempre se incluye para los componentes de numerador. Por ejemplo, si especifica 2 para el numerador, el modelo incluye los órdenes 2, 1 y 0. Si especifica 3 para el denominador, el modelo incluye los órdenes 3, 2 y 1. Las casillas de la columna Estacional solo se activan si se ha definido una periodicidad para el conjunto de datos activo.

Numerador. El orden de numerador de la función de transferencia especifica los valores previos de la serie independiente (predictora) seleccionada que se utilizan para predecir los valores actuales de la serie dependiente. Por ejemplo, un orden de numerador de 1 especifica que se utiliza el valor de una serie independiente de un período anterior (además del valor actual de la serie independiente) para predecir el valor actual de cada serie dependiente.

Denominador. El orden de denominador de la función de transferencia especifica cómo se utilizan las desviaciones respecto a la media de la serie para los valores previos de la serie independiente (predictora) seleccionada para predecir los valores actuales de la serie dependiente. Por ejemplo, un orden de denominador de 1 especifica que las desviaciones del valor medio de una serie independiente para un período de tiempo anterior se tienen en cuenta al predecir el valor actual de cada serie dependiente.

Diferencia. Especifica el orden de diferenciación aplicado a la serie independiente (predictora) seleccionada antes de estimar los modelos. La diferenciación es necesaria si hay tendencias y se utiliza para eliminar su efecto.

Órdenes estacionales. Los componentes estacionales de numerador, denominador y diferenciación tienen la misma función que los componentes no estacionales correspondientes. No obstante, en el caso de los órdenes estacionales, los valores de la serie actual se ven afectados por los valores de la serie anterior separados por uno o más períodos estacionales. Por ejemplo, para los datos mensuales (período estacional de 12), un orden estacional de 1 significa que el valor de la serie actual se ve afectado por el valor de la serie 12 períodos antes del actual. Un orden estacional de 1 para los datos mensuales equivale a la especificación de un orden no estacional de 12.

Retardo. Establecer un retardo provoca que la influencia del campo de entrada se retrase según el número de intervalos especificados. Por ejemplo, si el retardo se

establece en 5, el valor de la variable de entrada en el tiempo t no afecta a las predicciones hasta que han transcurrido cinco períodos (t + 5).

Transformación. La especificación de una función de transferencia para un conjunto de variables independientes también incluye una transformación opcional que se puede aplicar a dichas variables.

- *Ninguno*. No se lleva a cabo ninguna transformación.
- *Raíz cuadrada*. Se realiza una transformación de raíz cuadrada.
- *Log natural*. Se realiza una transformación logarítmica natural.

La pestaña *Valores atípicos* (Figura 10-13) ofrece varias opciones para tratar los valores atípicos de los datos. Las opciones son las siguientes:

No detectar valores atípicos ni modelarlos. Por defecto, los valores atípicos no se detectan ni modelan. Seleccione esta opción para desactivar la detección o el modelado de valores atípicos.

Detectar automáticamente los valores atípicos. Seleccione esta opción para realizar una detección automática de valores atípicos y seleccione uno o más de los tipos de valores atípicos que se muestran.

Tipos de valores atípicos que se detectarán. Seleccione los tipos de valores atípicos que desea detectar. Los tipos admitidos son:

- Aditivo (por defecto)
- Cambio de nivel (por defecto)
- Innovador
- Transitorio
- Aditivo estacional
- Tendencia local
- Parche aditivo

Resultados del modelo de serie temporal ajustado

Una vez ajustado el modelo de serie temporal ya hemos visto cómo se obtiene sobre la ruta el correspondiente icono del modelo (Figura 10-9). Si ahora hacemos

 ©Alfaomega-RC Libros

clic con el botón derecho del ratón en el icono del modelo *Series temporales* sobre la ruta, se obtiene la pantalla de resultados de la Figura 10-14. En la solapa *Modelo* de esta figura se observan los modelos estimados para cada serie y los estadísticos resumen. El modelizador automático ha hallado para la serie IRPF un modelo ARIMA(0,1,1)(0,1,0)$_{12}$ sin constante, para la serie del impuesto de sociedades IS00 un modelo ARIMA(0,1,0)(0,1,1)$_{12}$ sin constante y para la serie IVA0 un modelo ARIMA(0,1,1)(0,1,0)$_{12}$ sin constante. En las diferentes columnas de la Figura 10-14 se observan los parámetros de evaluación de los modelos seleccionados. Si en el campo *Ver* de la Figura 10-14 se elige *Avanzado* (Figura 10-15) se amplían los resultados sobre la estimación y la diagnosis de los modelos.

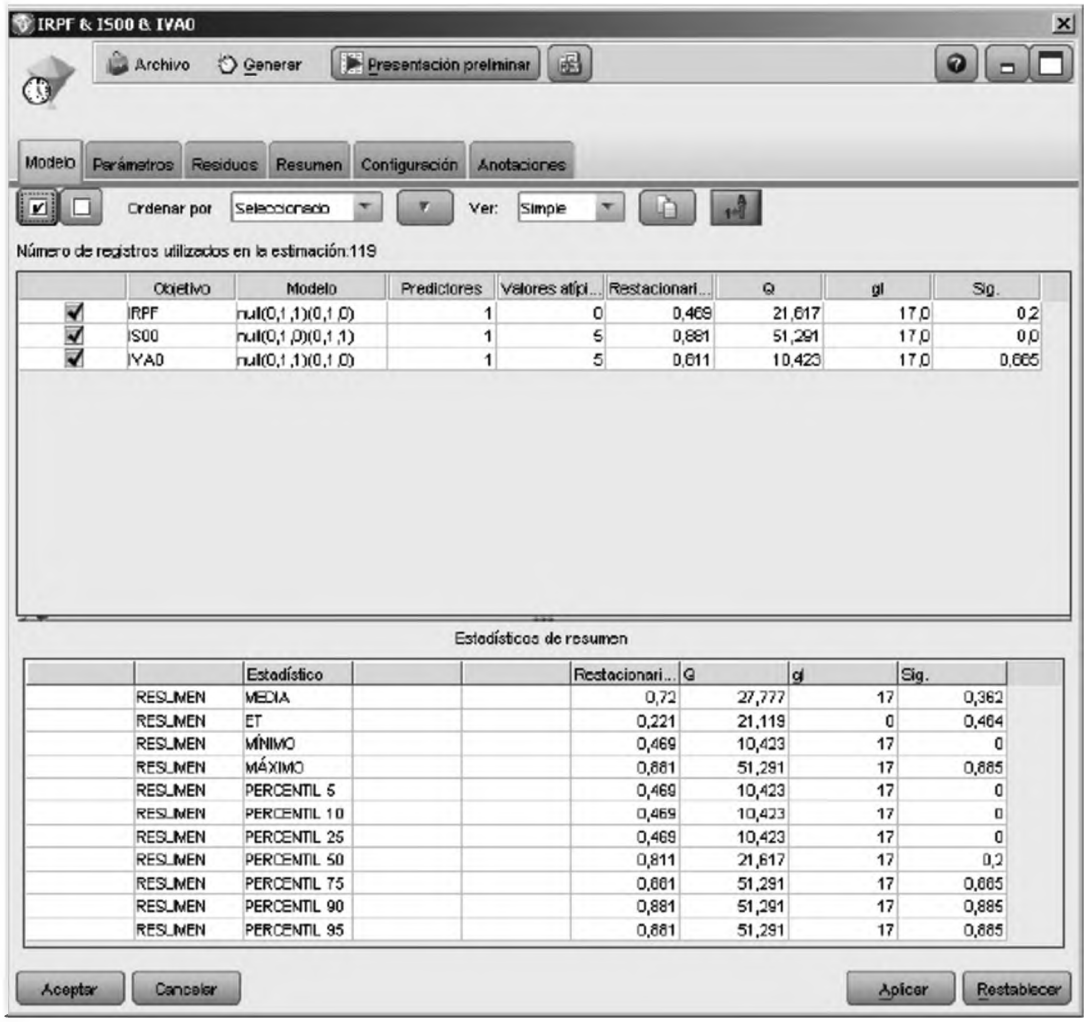


Figura 10-14

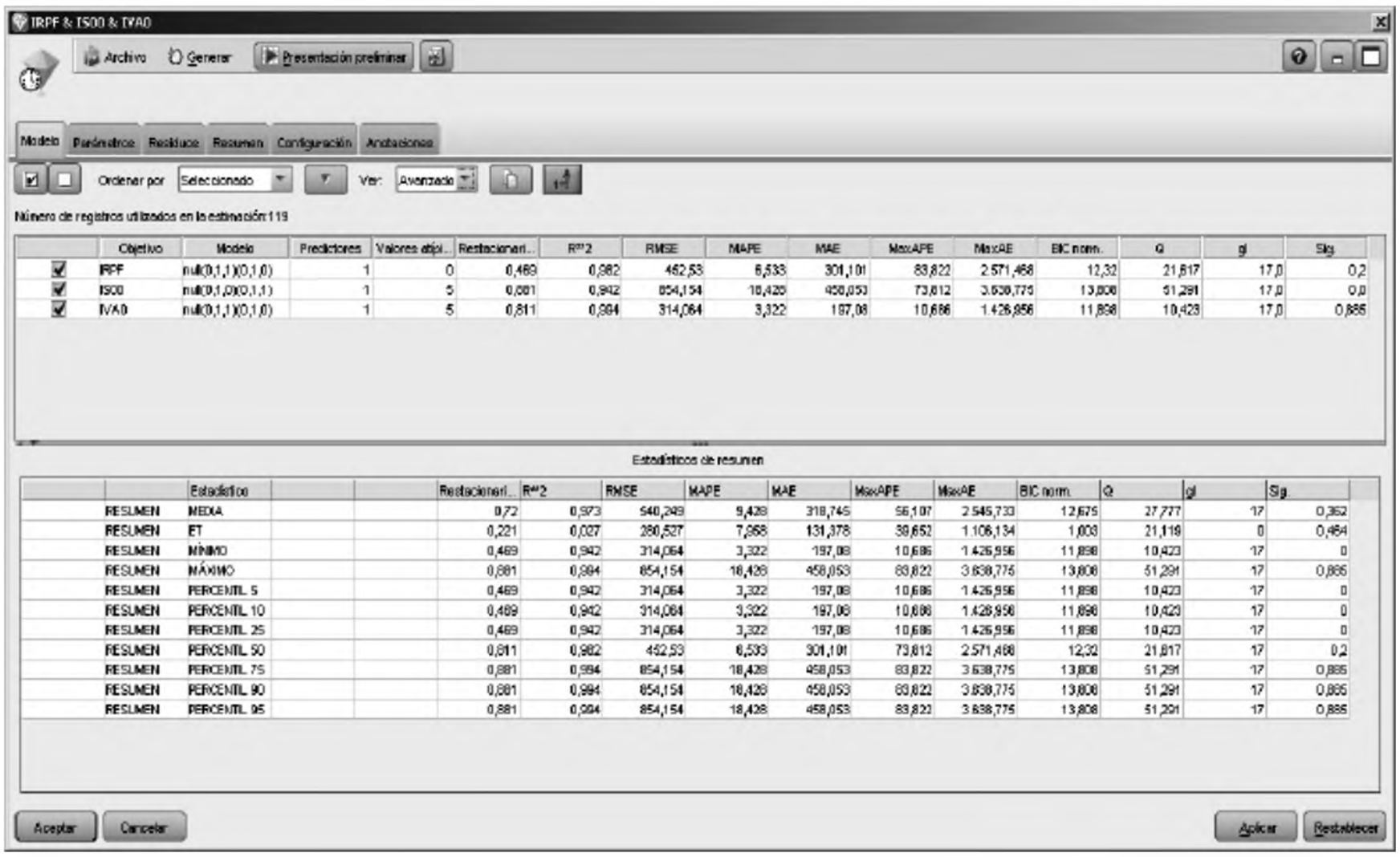


Figura 10-15

La solapa *Parámetros* de la pantalla de resultados del modelo presenta para cada serie temporal ajusatada (elegida en la pestaña *Mostrar parámetros del modelo*), los parámetros estimados y su diagnosis (Figuras 10-16, 10-17 y 10-18). Se observan p-valores muy pequeños en los modelos estimados para las tres series en estudio. Por lo tanto, los modelos ARIMA con transferencia estimados superan con éxito la diagnosis.

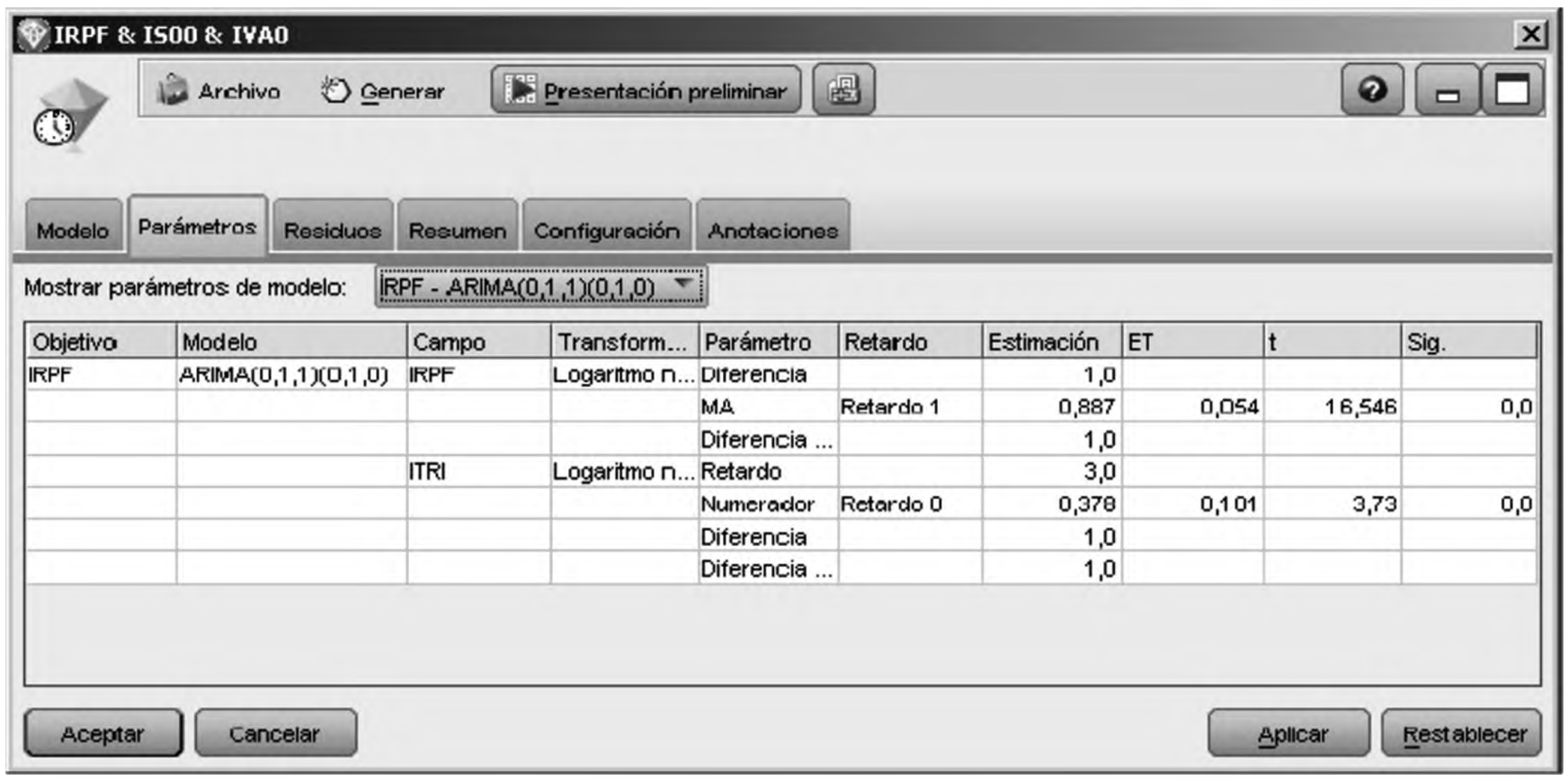


Figura 10-16

©Alfaomega-RC Libros

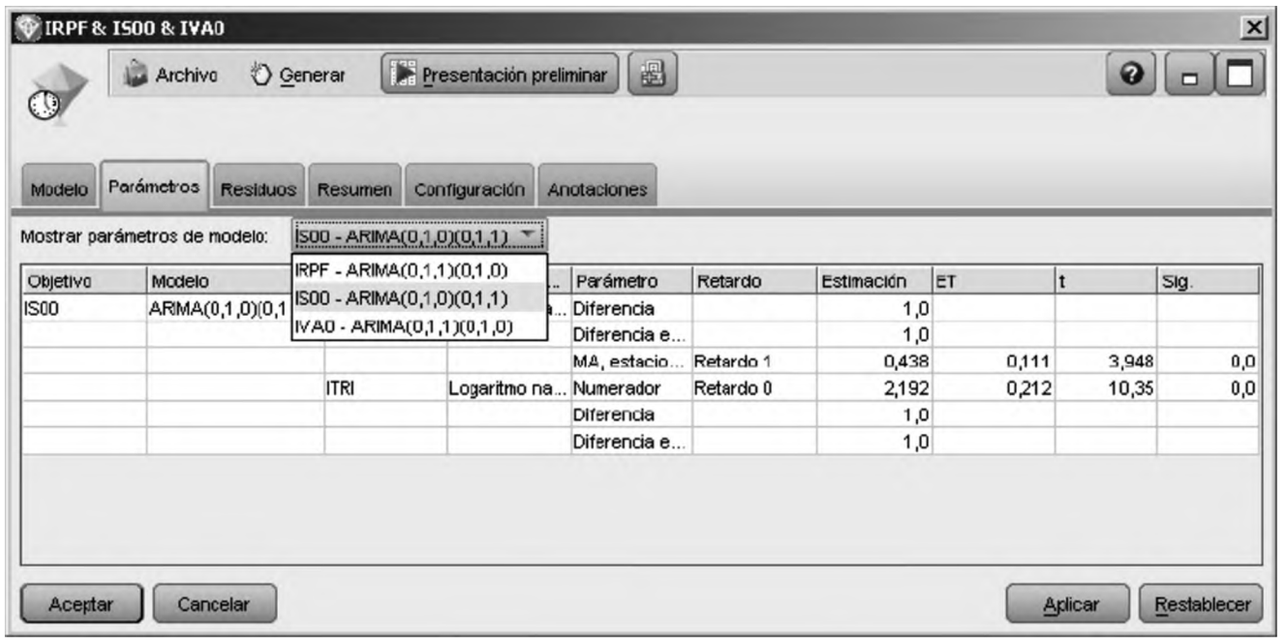


Figura 10-17

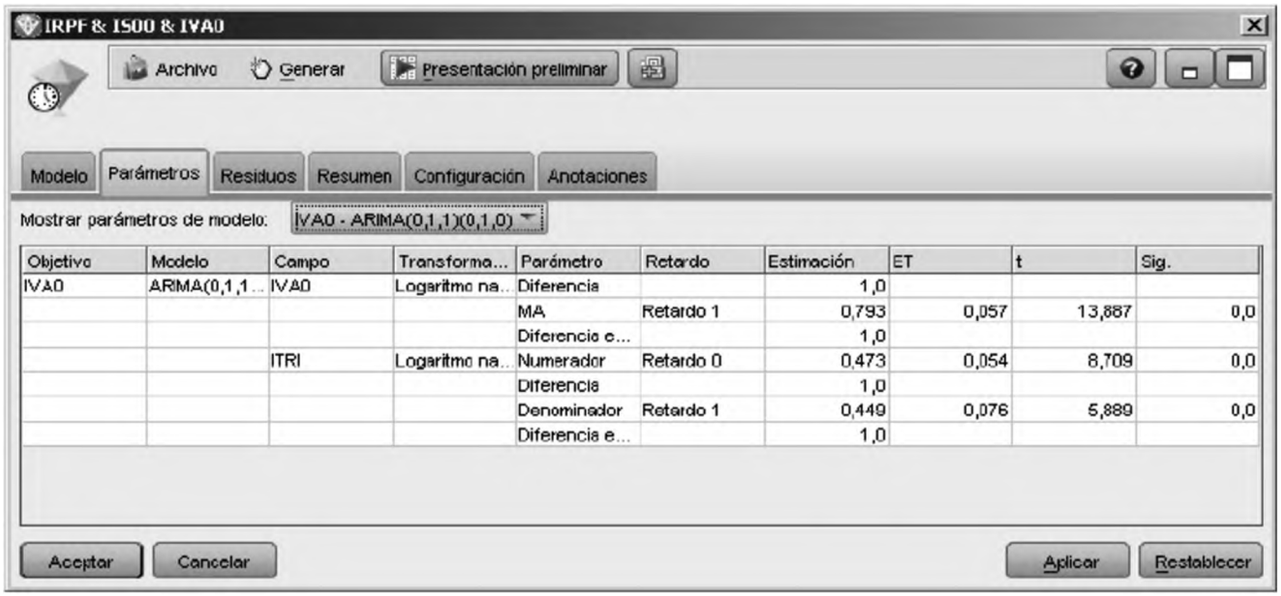


Figura 10-18

La solapa *Residuos* de la pantalla de resultados del modelo presenta para cada serie temporal ajustada (elegida en la pestaña *Mostrar gráfico de modelo*), los gráficos de las funciones de autocorrelación y autocorrelación parcial residuales con sus franjas de confianza al 95%, que se utilizan para realizar la diagnosis de los modelos ajustados (Figuras 10-19, 10-20 y 10-21). Para que el modelo sea correctamente diagnosticado es preciso que los términos de las funciones de autocorrelación no se salgan de las bandas de confianza. Se observa que prácticamente todos los términos están dentro de las bandas (salvo muy poquitos), lo que indica una diagnosis positiva que valida los modelos ajustados.

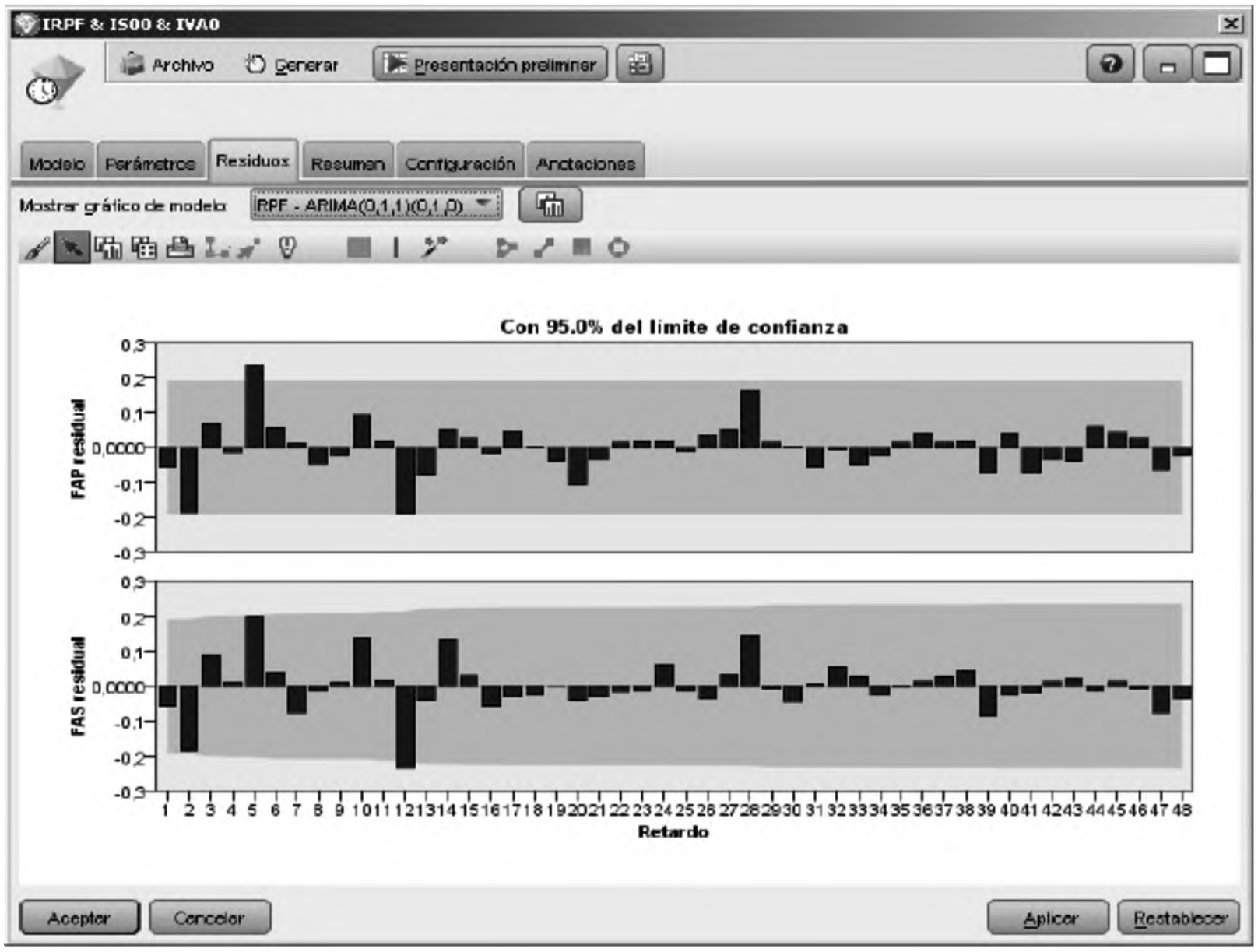


Figura 10-19

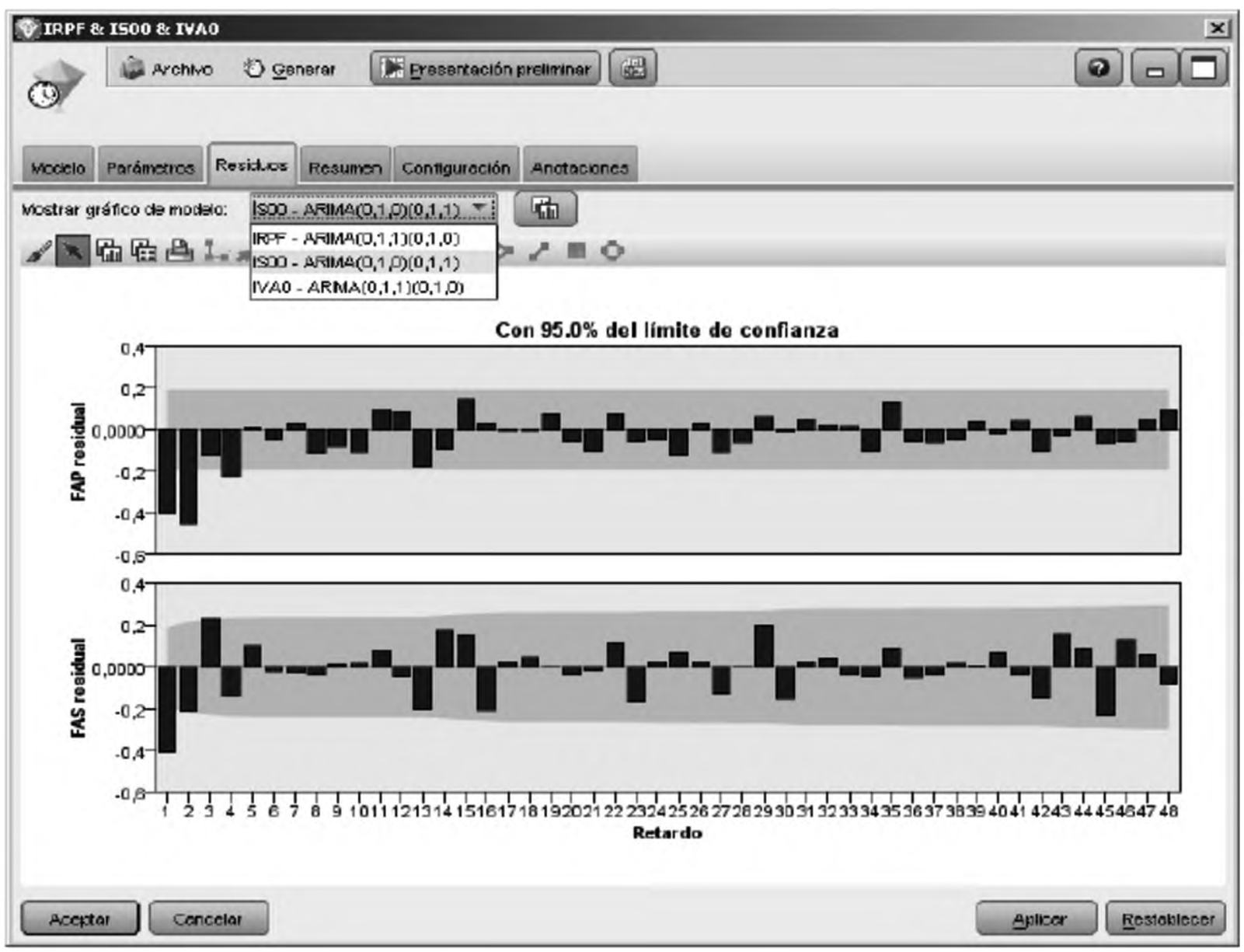


Figura 10-20

©Alfaomega-RC Libros

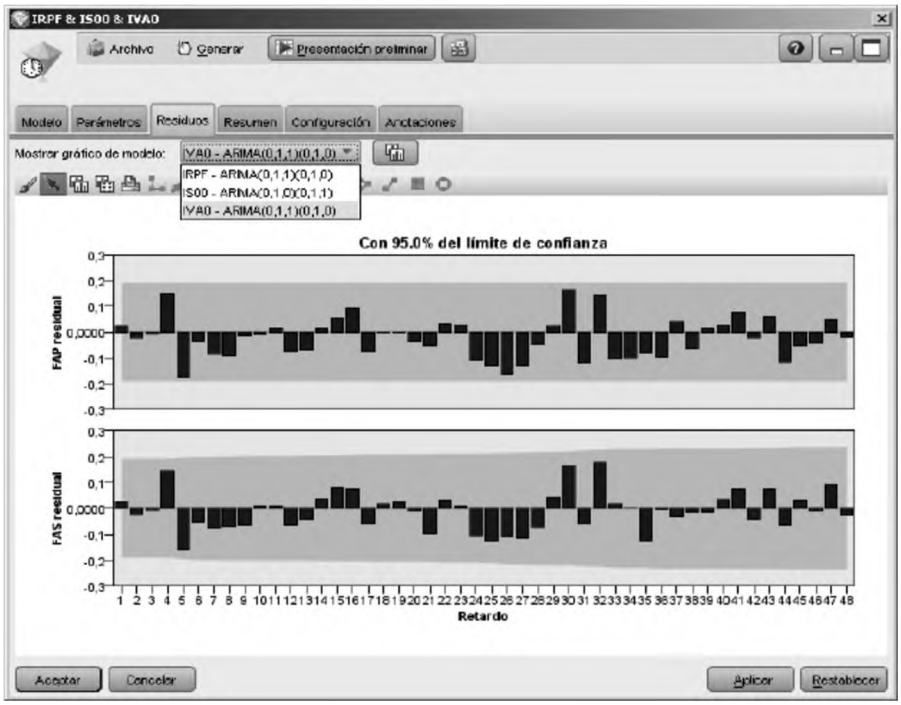


Figura 10-21

La solapa *Modelo* de la pantalla de resultados del modelo ajustado (Figura 10-14) presenta para cada modelo generado, los resultados por defecto del ajuste. En las sucesivas columnas de la pantalla se observa el tipo de modelo, el número de predictores especificados y la medida de bondad de ajuste (R cuadrado estacionaria es la medida por defecto). Si ha especificado métodos de valores atípicos, hay una columna donde aparece el número de valores atípicos detectados. Los resultados por defecto también incluyen columnas para Q de Ljung-Box, grados de libertad y valores de significación.

Si en el campo *Ver* de la Figura 10-14 se elige *Avanzado* (Figura 10-15) se amplían los resultados sobre la estimación y la diagnosis de los modelos, se muestran estas otras columnas:

- R cuadrado
- RMSE (raíz del error cuadrático promedio)
- MAPE (error absoluto porcentual promedio)
- MAE (error absoluto promedio)

• MaxAPE (error absoluto máximo porcentual)

• MaxAE (error absoluto máximo)

• Norm. BIC (criterio de información bayesiano normalizado)

El botón *Generar* de la solapa *Modelo* permite volver a generar un nodo de modelado *Serie temporal* en la ruta. Tenemos las opciones:

• *Generar nodo de modelado*. Coloca un nodo de modelado Serie temporal en una ruta con la configuración usada para crear este conjunto de modelos. Esto resultará útil, por ejemplo, si tiene una ruta en la que desee usar esta configuración de modelo y ya no tiene el nodo de modelado que se utilizó para generarla.

• *Modelo a paleta*. Sitúa un nugget de modelo con todos los objetivos en el administrador de modelos.

En la solapa *Modelo* tenemos también las siguientes opciones:

Casillas de verificación. Seleccione los modelos que desea utilizar en la puntuación. Todas las casillas están activadas por defecto. Los botones *Seleccionar todos* y *Anular la selección de todos* afectan a todas las casillas en una única operación.

Ordenar por. Le permite ordenar las filas de resultados en orden ascendente o descendente de una columna especificada de la presentación. La opción "Seleccionado" ordena el resultado según una o más filas seleccionadas por casillas de verificación. Esto resultaría útil, por ejemplo, para forzar que los campos objetivo denominados "Mercado_1" a "Mercado_9" aparezcan antes de "Mercado_10", ya que el orden de clasificación por defecto muestra "Mercado_10" justo después de "Mercado_1."

Ver. La vista por defecto (simple) muestra el conjunto básico de columnas de resultados. La opción *Avanzado* muestra columnas adicionales para medidas de bondad de ajuste.

Número de registros utilizados en la estimación. Número de filas existentes en el archivo de datos fuente original.

©Alfaomega-RC Libros

Destino. Campo o campos identificados como campos objetivo (aquellos con un papel de Objetivo) en el nodo Tipo.

Modelo. Tipo de modelo usado para este campo objetivo.

Predictores. El número de predictores (aquellos con un papel de Entrada) utilizados para este campo objetivo.

Valores atípicos. Esta columna solo aparece si ha solicitado (en el modelizador experto o los criterios ARIMA) la detección automática de valores atípicos. El valor mostrado es el número de valores atípicos detectados.

R-cuadrado estacionaria. Una medida que compara la parte estacionaria del modelo con un modelo de promedio simple. Esta medida es preferible al R-cuadrado ordinario cuando existe tendencia o patrón estacional. R-cuadrado puede ser negativa con un rango desde menos infinito hasta 1. Los valores negativos significan que el modelo estudiado es peor que el modelo basal. Los valores positivos significan que el modelo estudiado es mejor que el modelo basal.

R cuadrado. Medida de la bondad de ajuste de un modelo lineal; en ocasiones recibe el nombre de coeficiente de determinación. Es la proporción de la variación de la variable dependiente explicada por el modelo de regresión. Puede tomar un valor entre 0 y 1. Un valor pequeño indica que el modelo no se ajusta bien a los datos.

RMSE. Raíz del error cuadrático promedio. La raíz cuadrada del error cuadrático promedio. Una medida de cuánto se desvía la serie dependiente del nivel pronosticado por el modelo, expresado en las mismas unidades que la serie dependiente.

MAPE. Error absoluto porcentual promedio. Medida de la desviación de la serie dependiente del nivel pronosticado por el modelo. Es independiente de las unidades utilizadas y se puede utilizar para comparar series con distintas unidades.

MAE. Error absoluto promedio. Mide la desviación de la serie del nivel pronosticado por el modelo. El MAE se informa en las unidades originales de la serie.

MaxAPE. Error absoluto máximo porcentual. El mayor error pronosticado, expresado como porcentaje. Esta medida es útil para imaginar el peor escenario de un caso en las predicciones.

MaxAE. Error absoluto máximo. El mayor error pronosticado, expresado en las mismas unidades que la variable dependiente. Al igual que el MaxAPE, es útil para imaginar el peor escenario de los casos en la predicción. El error absoluto máximo y el error absoluto máximo porcentual pueden darse en distintos puntos de la serie. Por ejemplo, si el error absoluto de un valor de la serie grande es ligeramente mayor que el error absoluto de un valor de la serie pequeño. En ese caso el error absoluto máximo se obtendrá en el valor de la serie mayor y el error absoluto máximo porcentual corresponderá al valor de la serie menor.

BIC normalizado. Criterio de información bayesiano normalizado. Una medida general del ajuste global del modelo que intenta tener en cuenta la complejidad del modelo. Es una medida basada en el error cuadrático promedio que incluye una penalización para el número de parámetros presentes en el modelo y la longitud de la serie. La penalización elimina la ventaja de los modelos con mayor número de parámetros, haciendo que el estadístico sea fácil de comparar entre distintos modelos para la misma serie.

Q. Estadístico Q de Ljung-Box. Prueba de la aleatoriedad de los errores residuales de este modelo.

gl. Grados de libertad. Número de parámetros del modelo que pueden variar al calcular un objetivo específico.

Sig. Valor de significación del estadístico de Ljung-Box. Un valor de significación inferior a 0,05 indica que los errores residuales no son aleatorios.

Estadísticos de resumen. Esta sección contiene diferentes estadísticos de resumen para las distintas columnas, incluidos los valores de la media, mínimo, máximo y los percentiles.

La pestaña *Parámetros* de la pantalla de resultados del ajuste de una serie temporal (Figuras 10-16, 10-17 y 10-18) enumera los detalles de los distintos parámetros que se utilizaron para crear un modelo seleccionado. Tenemos las opciones siguientes:

Mostrar parámetros de modelo. Seleccione el modelo para el que desea mostrar los detalles de parámetros.

 ©Alfaomega-RC Libros

Destino. Nombre del campo objetivo (con el papel Objetivo) pronosticado por este modelo.

Modelo. Tipo de modelo usado para este campo objetivo.

Campo (solo modelos ARIMA). Contiene una entrada para cada una de las variables utilizadas en el modelo, con el objetivo en primer lugar, seguido por los predictores, si los hubiera.

Transformación. Indica qué tipo de transformación se ha especificado para este campo antes de generar el modelo, si hay alguna transformación.

Parámetro. Parámetro del modelo para el que se muestran los siguientes detalles:

• *Retardo* (solo modelos ARIMA). Indica los retardos, en caso de que los haya, que se tienen en cuenta para este parámetro en el modelo.

• *Estimación*. Estimación del parámetro. Este valor se utiliza para calcular el valor pronosticado y los intervalos de confianza para el campo objetivo.

• *SE*. Error estándar de la estimación del parámetro.

• *t*. Valor de la estimación del parámetro dividido entre el error estándar.

• *Sig*. Nivel de significación para la estimación del parámetro. Los valores por encima de 0,05 no se consideran estadísticamente significativos.

La pestaña *Residuos* (Figuras 10-19, 10-20 y 10-21) muestra la función de autocorrelación (FAS) y la función de autocorrelación parcial (FAP) de los residuos (diferencia entre los valores esperados y los reales) de cada modelo generado. En la opción *Mostrar gráfico de modelo* se selecciona el modelo cuyas FAS y FAP de residuos se desea mostrar.

La ficha *Resumen* de la pantalla de resultados del modelo de serie temporal ajustado muestra información sobre el propio modelo (*Análisis*), los campos utilizados en el modelo (*Campos*), la configuración utilizada al generar el modelo (*Configuración de creación*) y el entrenamiento del modelo (*Resumen de entrenamiento*). Cuando se examina el nodo por primera vez, los resultados de la ficha *Resumen* aparecen contraídos. Para ver los resultados de interés, utilice el control de expansión situado a la izquierda de un elemento con objeto de

desplegarlo, o bien pulse en el botón *Expandir todo* para mostrar todos los resultados. Para ocultar los resultados cuando haya terminado de consultarlos, utilice el control de expansión con objeto de contraer los resultados específicos que desee ocultar o pulse en el botón *Contraer todo* para contraer todos los resultados. Tenemos lo siguiente:

Análisis. Muestra información sobre el modelo específico.

Campos. Enumera los campos utilizados como objetivo y entradas en la generación del modelo.

Configuración de creación. Contiene información sobre la configuración que se utiliza en la generación del modelo.

Resumen de entrenamiento. Muestra el tipo del modelo, la ruta utilizada para crearlo, el usuario que lo creó, cuándo se generó y el tiempo que se tardó en generar el modelo.

La pestaña *Configuración* de la pantalla de resultados del ajuste del modelo de serie temporal le permite especificar los campos adicionales que se van a crear mediante la operación de modelado. Las opciones son las siguientes:

Crear nuevos campos para cada modelo que se puntuará. Permite especificar los nuevos campos que se crearán para cada modelo que se puntuará. Se tiene:

• *Calcular límites de confianza superior e inferior*. Si está seleccionada, se crearán nuevos campos (con los prefijos por defecto $TSLCI- y $TSUCI-) para los intervalos de confianza superior e inferior, respectivamente, de cada campo objetivo, además de los totales de estos valores.

• *Calcular residuos de ruido*. Si está activada, creará un nuevo campo (con el prefijo por defecto $TSNR-) para los residuos del modelo de cada campo objetivo, junto con un total de estos valores.

La operación de modelado de series temporales crea varios campos nuevos con el prefijo $TS- como se indica a continuación:

$TS-nombrecol. Valor pronosticado por el modelo para cada serie objetivo.

©Alfaomega-RC Libros

$TSLCI-nombrecol.	Los intervalos de confianza más bajos para cada serie pronosticada.*
$TSUCI-nombrecol.	Los intervalos de confianza más altos para cada serie pronosticada.*
$TSNR-nombrecol.	Valor de residuo de ruido para cada columna de datos del modelo generado.*
$TS-Total	Total de los valores de $TS-nombrecol de esta fila.
$TSLCI-Total	Total de los valores de $TSLCI-nombrecol de esta fila.*
$TSUCI-Total	Total de los valores de $TSUCI-nombrecol de esta fila.*
$TSNR-Total	Total de los valores de $TSNR-nombrecol de esta fila.*

* La visibilidad de estos campos (por ejemplo, en los resultados del nodo *Tabla* conectado) depende de las opciones de la pestaña *Configuración* de la pantalla de resultados del modelo Serie temporal.

Generación de modelos de series temporales

Esta sección proporciona información general acerca de algunos aspectos de generación de modelos de series temporales:

- Generación de varios modelos
- Uso de los modelos de series temporales en predicciones
- Nueva estimación y predicción

Generación de varios modelos

El modelado de series temporales en IBM® SPSS® Modeler genera un único modelo (ARIMA o de suavizado exponencial) para cada campo objetivo. De esta manera, si tiene varios campos objetivo, SPSS Modeler genera varios modelos en una única operación, lo que le permitirá ahorrar tiempo y comparar la configuración de cada modelo.

Si desea comparar un modelo ARIMA y un modelo de suavizado exponencial para el mismo campo objetivo, puede realizar diferentes ejecuciones del nodo Serie temporal especificando de un modelo distinto cada vez.

Uso de los modelos de series temporales en predicciones

Una operación de generación de series temporales usa una serie concreta de casos ordenados, conocida como amplitud de estimación, para generar un modelo

que pueda utilizarse para predecir valores futuros de las series. Este modelo contiene información acerca del período de tiempo usado, incluido el intervalo. Para poder realizar predicciones con este modelo, debe utilizarse la misma información de período temporal e intervalo con la misma serie para la variable objetivo y las variables predictoras.

Por ejemplo, supongamos que a principios de enero desea pronosticar las ventas mensuales del Producto 1 durante el primer trimestre del año. Para ello, genera un modelo utilizando los datos reales de ventas mensuales para el Producto 1 de enero a diciembre del año anterior (que denominaremos Año 1), configurando Intervalo de tiempo como "Meses". A continuación, se puede utilizar el modelo para predecir las ventas del Producto 1 durante el primer trimestre del Año 2.

De hecho, puede pronosticar cualquier número de meses del futuro pero, como es lógico, cuanto más lejos en el tiempo se encuentren las predicciones, menos eficaz será el modelo. No obstante, no será posible predecir las tres primeras semanas del Año 2, ya que el intervalo usado para generar el modelo era "Meses". Tampoco tendría sentido usar este modelo para predecir las ventas del Producto 2, ya que un modelo de series temporales solo es relevante para los datos que se usaron para definirlo.

Nueva estimación y predicción

El período de estimación está codificado internamente en el modelo que se ha generado. Es decir, se ignorarán los valores situados fuera del período de estimación si aplica el modelo actual a datos nuevos. De esta forma, un modelo de serie temporal se debe volver a estimar cada vez que haya nuevos datos disponibles, en contraste con otros modelos de IBM® SPSS® Modeler, que se pueden volver a aplicar para obtener puntuaciones sin necesidad de modificarlos.

Para seguir con el ejemplo anterior, supongamos que, para principios de abril en el Año 2, tiene los datos de ventas mensuales reales para el período comprendido entre enero y marzo. No obstante, si vuelve a aplicar el modelo que generó a principios de enero, volverá a predecir de enero a marzo e ignorará los datos de ventas conocidos para ese período.

 ©Alfaomega-RC Libros

La solución es generar un nuevo modelo basado en los datos reales actualizados. Suponiendo que no cambia los parámetros de predicción, el nuevo modelo se puede utilizar para predecir los tres meses siguientes (de abril a junio). Si aún puede acceder a la ruta que se utilizó para generar el modelo original, puede simplemente sustituir la referencia al archivo fuente de esa ruta con una referencia al archivo que contiene los datos actualizados y volver a ejecutar la ruta para generar el nuevo modelo. No obstante, si solo tiene el modelo original guardado en un archivo, puede seguir utilizándolo para generar un nodo Serie temporal que podrá añadir después a una nueva ruta con una referencia al archivo fuente actualizado. La ejecución de esta nueva ruta generará, a continuación, el nuevo modelo necesario, siempre que en esta nueva ruta antes del nodo Serie temporal haya un nodo Intervalos de tiempo en el que el intervalo esté configurado como "Meses".

CAPÍTULO 11

PRONÓSTICOS Y TENDENCIAS CON HERRAMIENTAS DE SAS

PREDICCIONES INCONDICIONALES DETERMINISTAS Y SUAVIZADO CON EL MODELIZADOR DE SAS

En la ruta *Soluciones* → *Análisis* (Figura 11-1), SAS dispone de un modelizador automático que permite ajustar series a modelos incondicionales deterministas (y de todos los tipos), suavizarlas y obtener predicciones.

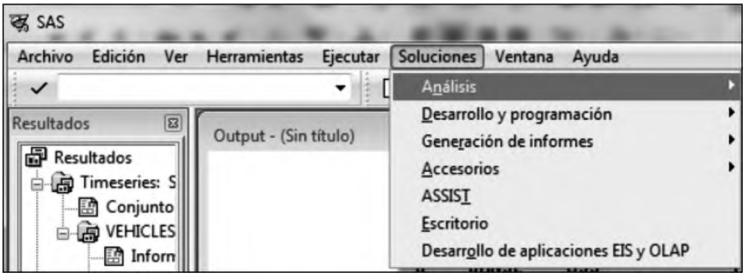


Figura 11-1

Se accede al modelizador mediante *Soluciones* → *Análisis* → *Sistema de predicción de series temporales* (Figura 11-2). Se obtiene la Figura 11-3 en la que hay que elegir el conjunto de datos que contiene la serie temporal a analizar *Data Set*, la variable que contiene la secuencia temporal *Time ID* y el intervalo estacional *Interval*. Ya hemos visto anteriormente en este capítulo que la serie es estacional mensual. Para ello se hace clic en *Browse* a la derecha de *Data Set* y se elige librería, conjunto de datos, variable Id e intervalo estacional (Figura 11-4). Se hace clic en OK y ya se memorizan las características elegidas (Figura 11-5). Se hace clic en *Produce Forecast* y a la derecha del campo *Series to Process* de la Figura 11-6 se hace clic en *Select* para elegir la variable para la que hay que realizar las predicciones de entre la lista de todas las variables del fichero.

A continuación se hace clic con el botón derecho del ratón sobre cualquier zona de la pantalla y en el menú emergente resultante se elige *Options*. Se despliega el menú de la Figura 11-7 cuyas opciones permitirán administrar el modelizador.

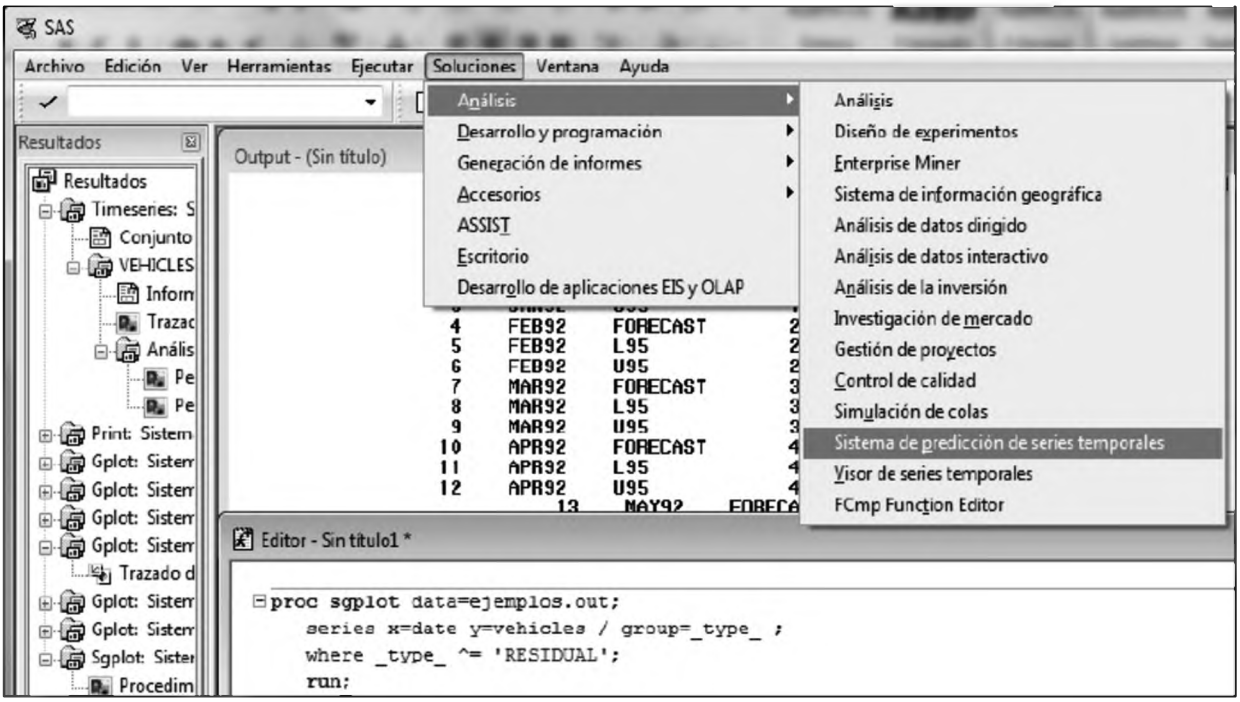


Figura 11-2

©Alfaomega-RC Libros

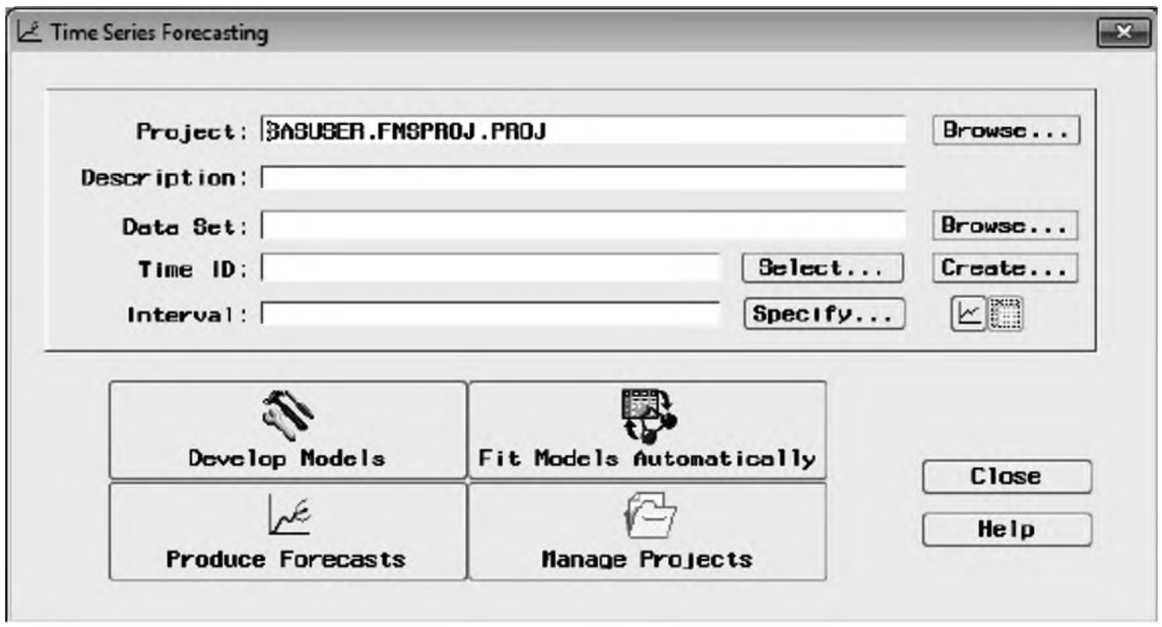


Figura 11-3

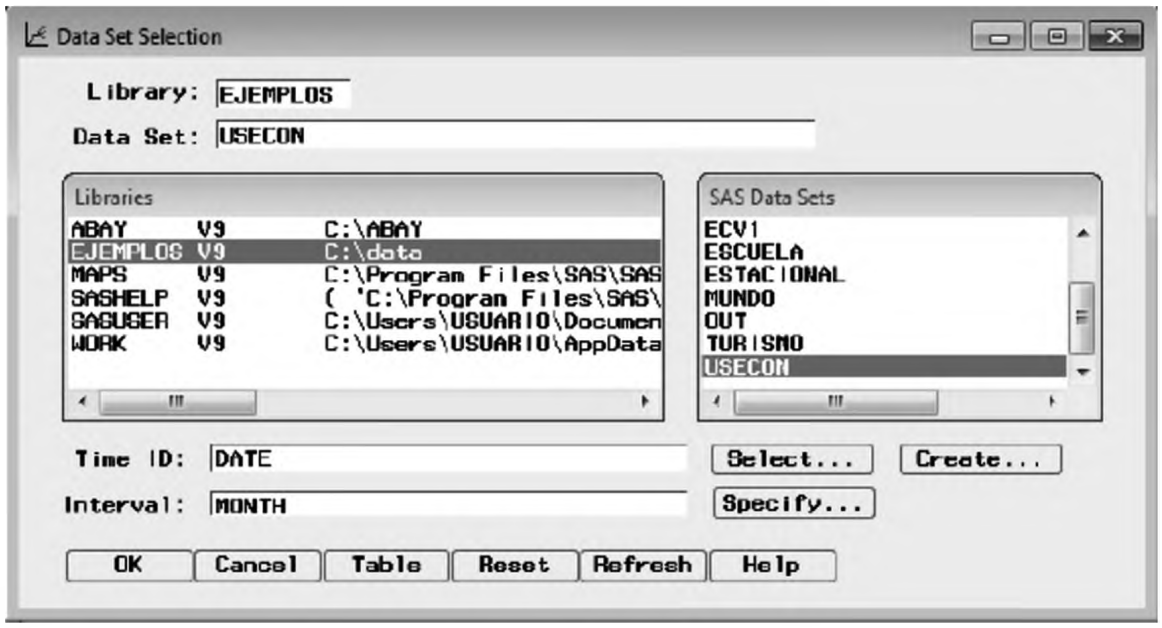


Figura 11-4

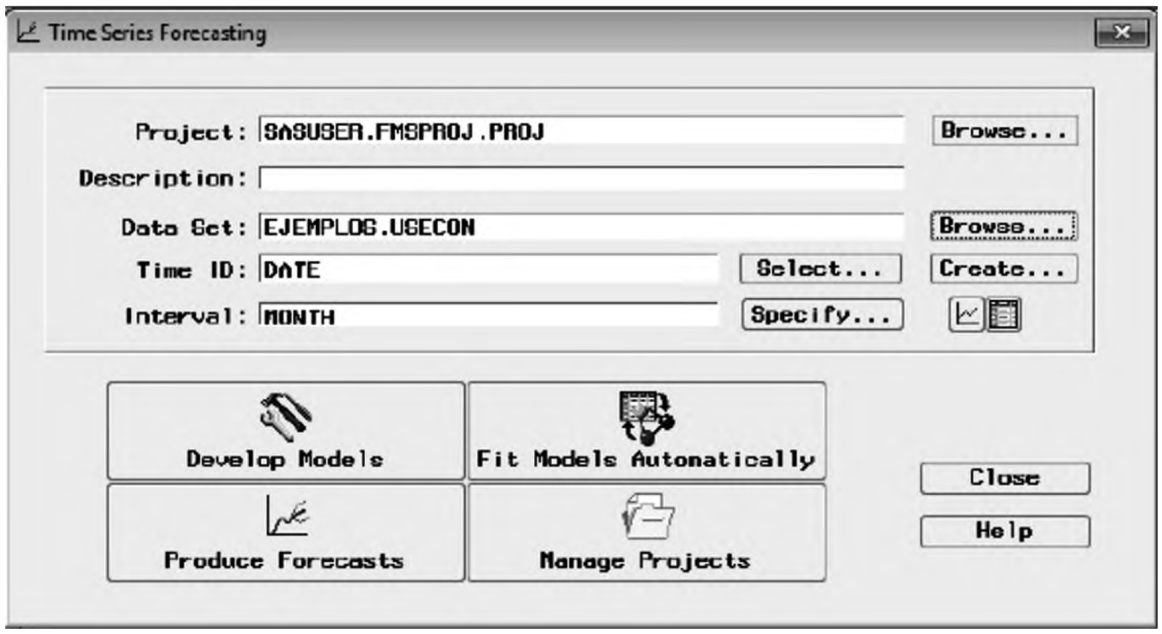


Figura 11-5

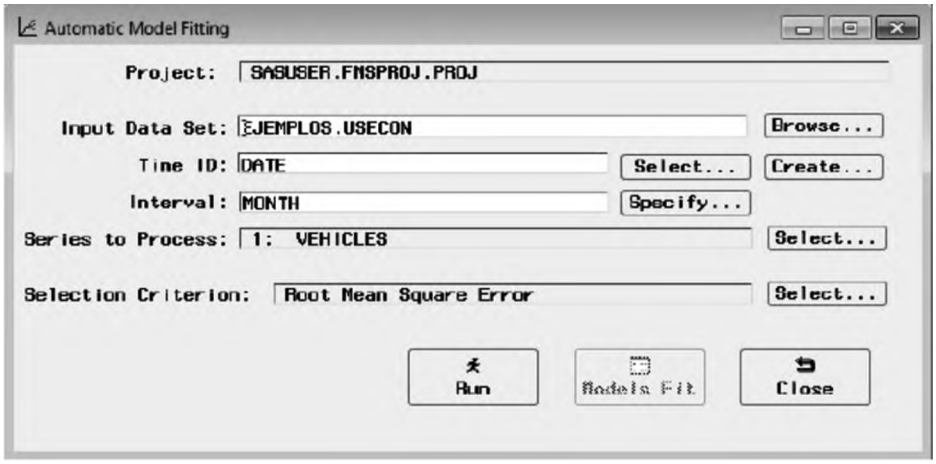


Figura 11-6

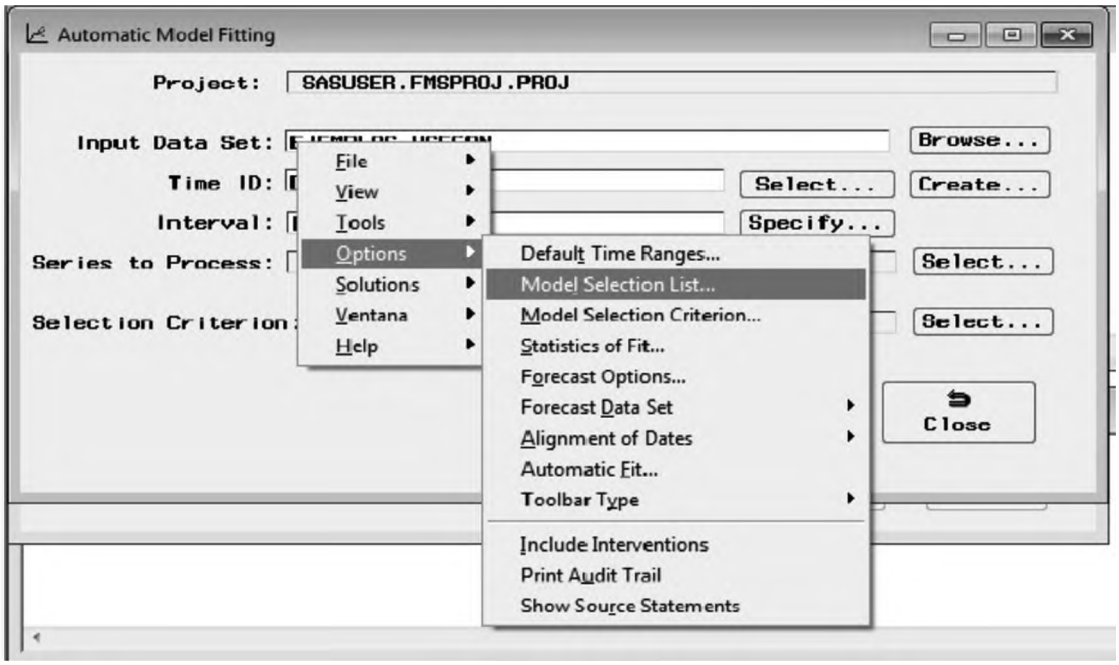


Figura 11-7

La opción *Default Time Ranges* de la Figura 11-7 permite fijar el horizonte de predicción y definir el período de ajuste (Figura 11-8). La opción *Model Selection List* permite elegir los tipos de modelos para suavizar la serie y realizar las predicciones (Figura 11-9). En nuestro caso elegimos los posibles modelos incondicionales deterministas para que el modelizador elija el mejor. La opción *Statistics Fit selection* permite elegir los estadísticos de ajuste para la elección del mejor modelo de entre los propuestos (Figura 11-10). La opción *Model Selection Criterion* presenta los criterios seleccionados para la elección del mejor modelo de ajuste (Figura 11-11). La opción *Forecast Options* permite elegir los límites de confianza para las predicciones (Figura 11-12).

©Alfaomega-RC Libros

Elegidas todas estas características, se hace clic en *Run* y se obtiene que el mejor modelo es el suavizado estacional de Winters multiplicativo (Figura 11-13). Seleccionado este modelo, se hace clic en *Graph* y se obtiene la gráfica de la serie (Figura 11-14).

Los iconos de la franja vertical de la derecha de esta figura permiten ver la gráfica de la serie (Figura 11-14), ver la gráfica de los residuos (Figura 11-15), ver la gráfica de la serie original, la serie suavizada y las predicciones en el mismo gráfico (Figura 11-16) y ver una tabla con los valores originales de la serie, los valores predichos, los errores y los intervalos de confianza para las predicciones (Figura 11-17).

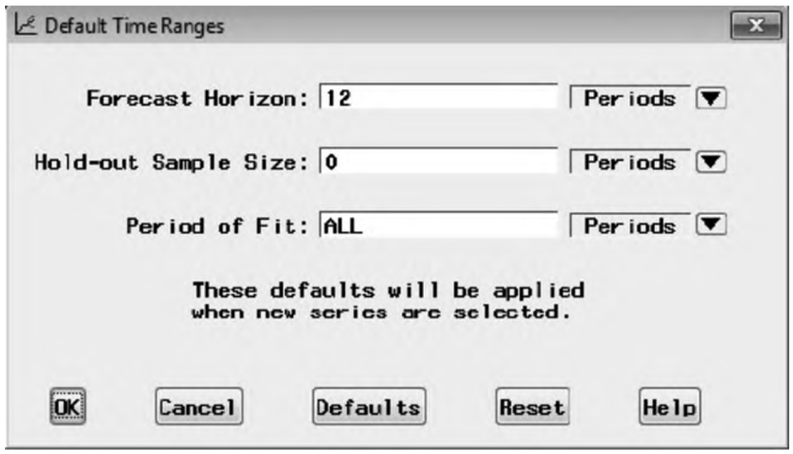


Figura 11-8

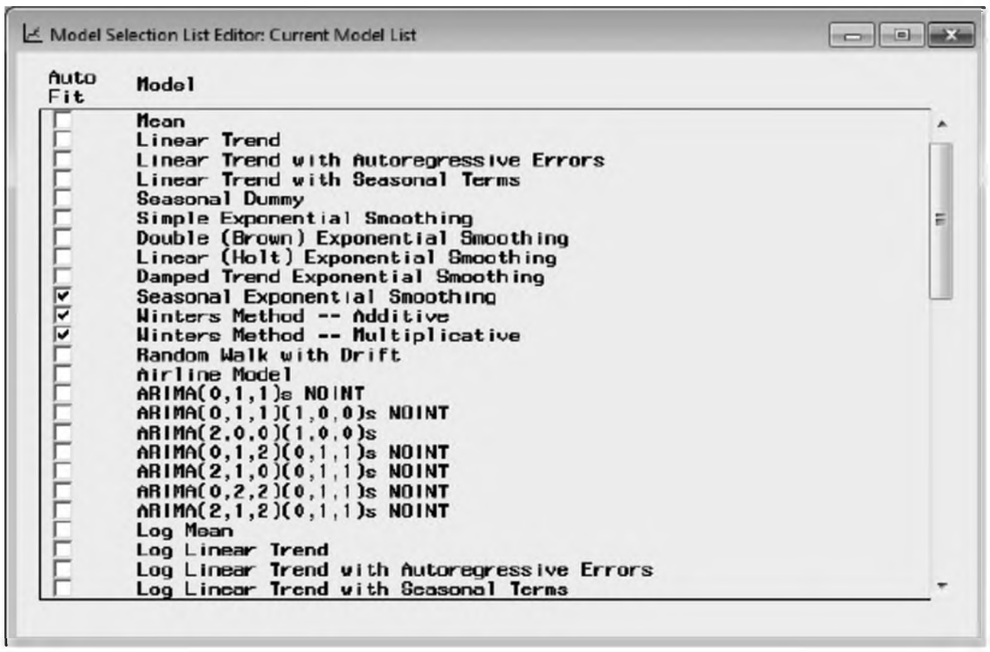


Figura 11-9

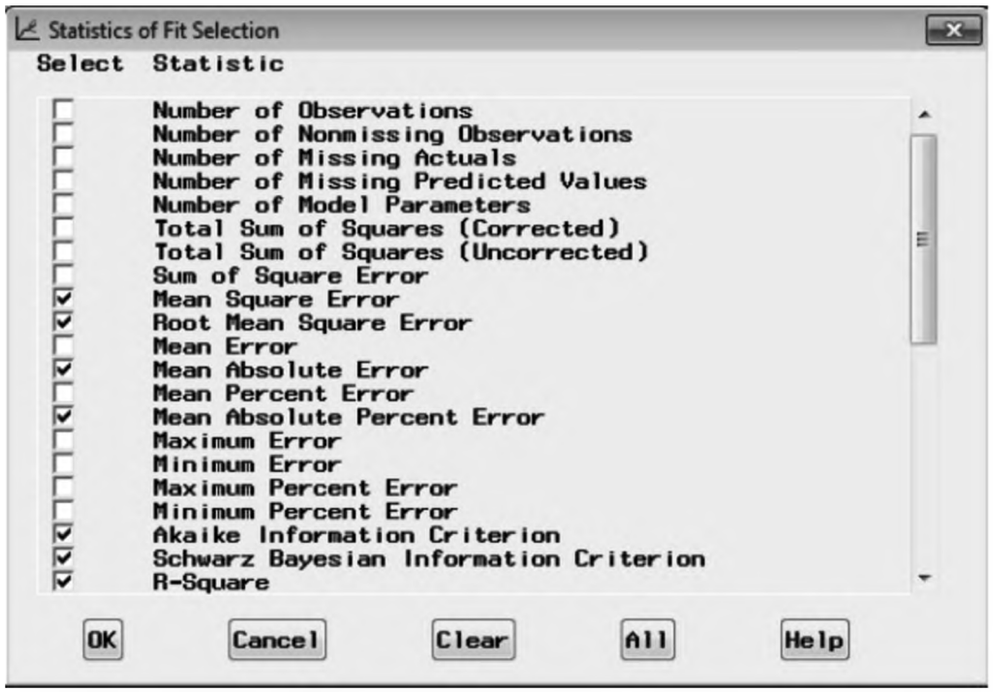


Figura 11-10

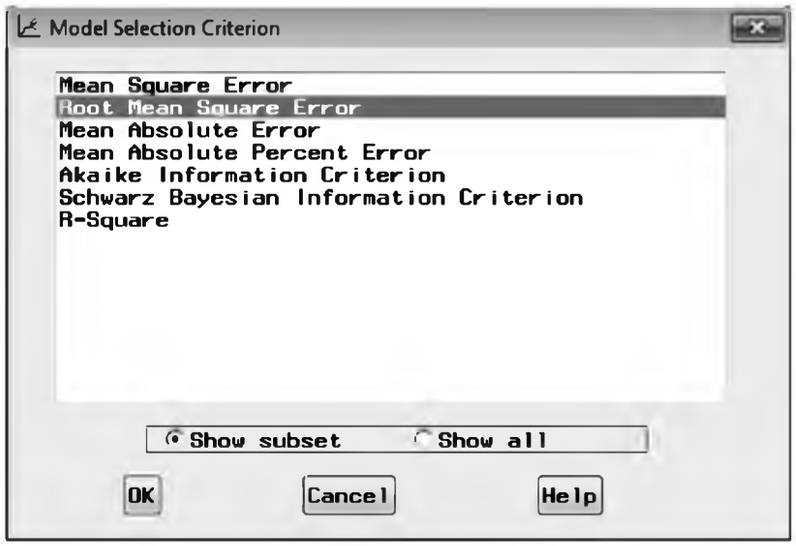


Figura 11-11

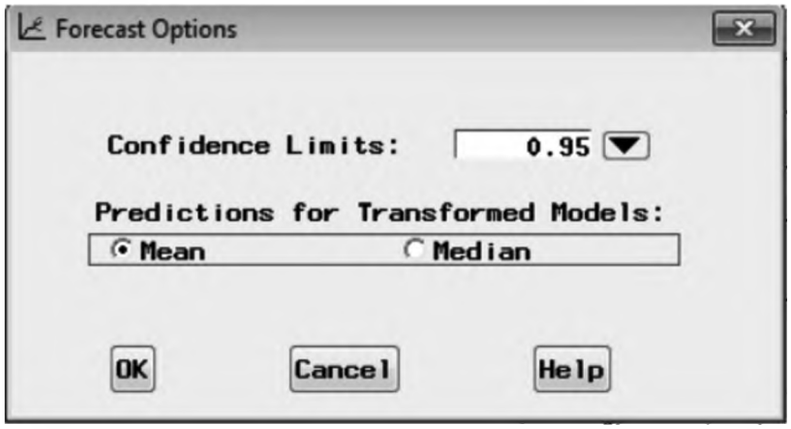


Figura 11-12

©Alfaomega-RC Libros

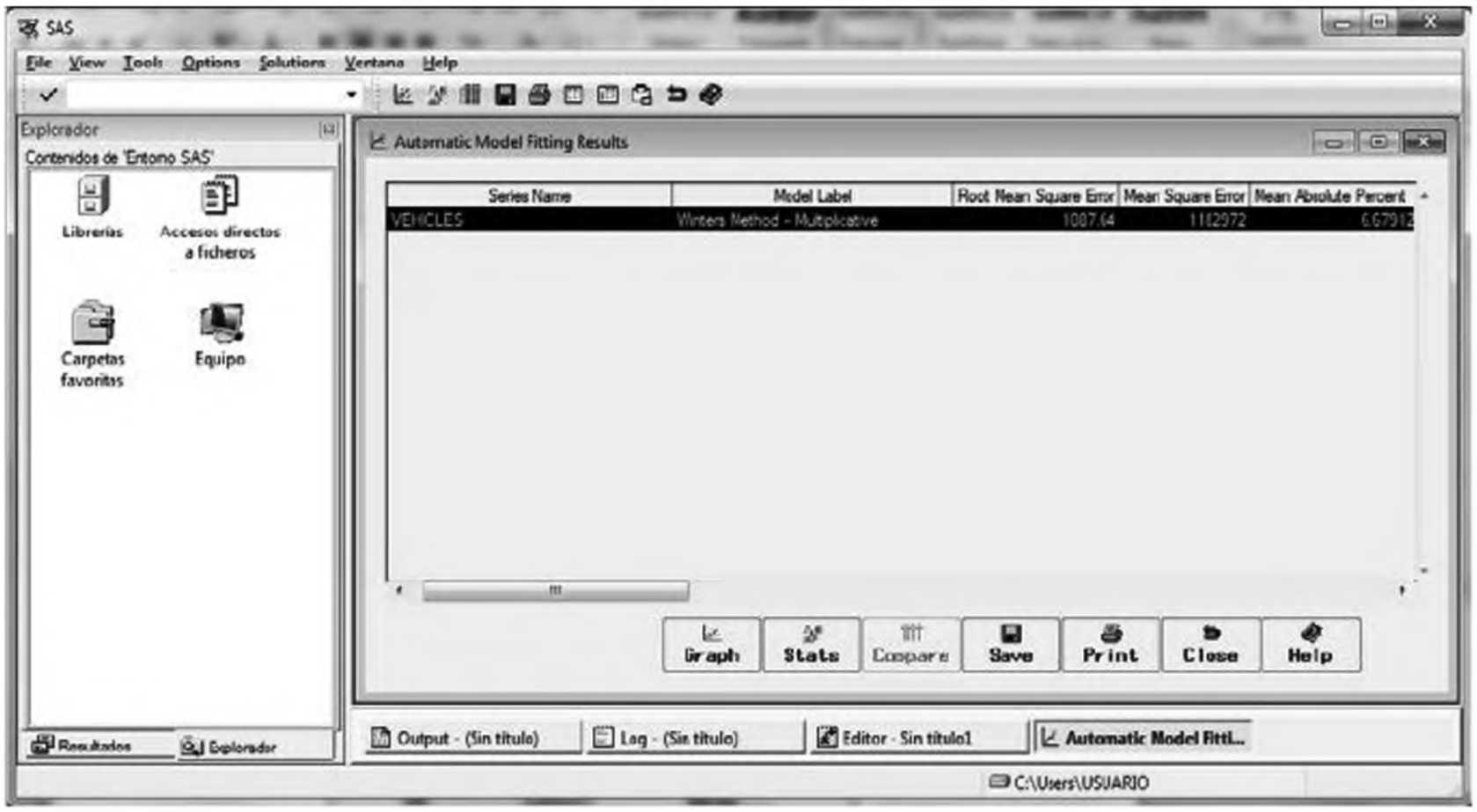


Figura 11-13

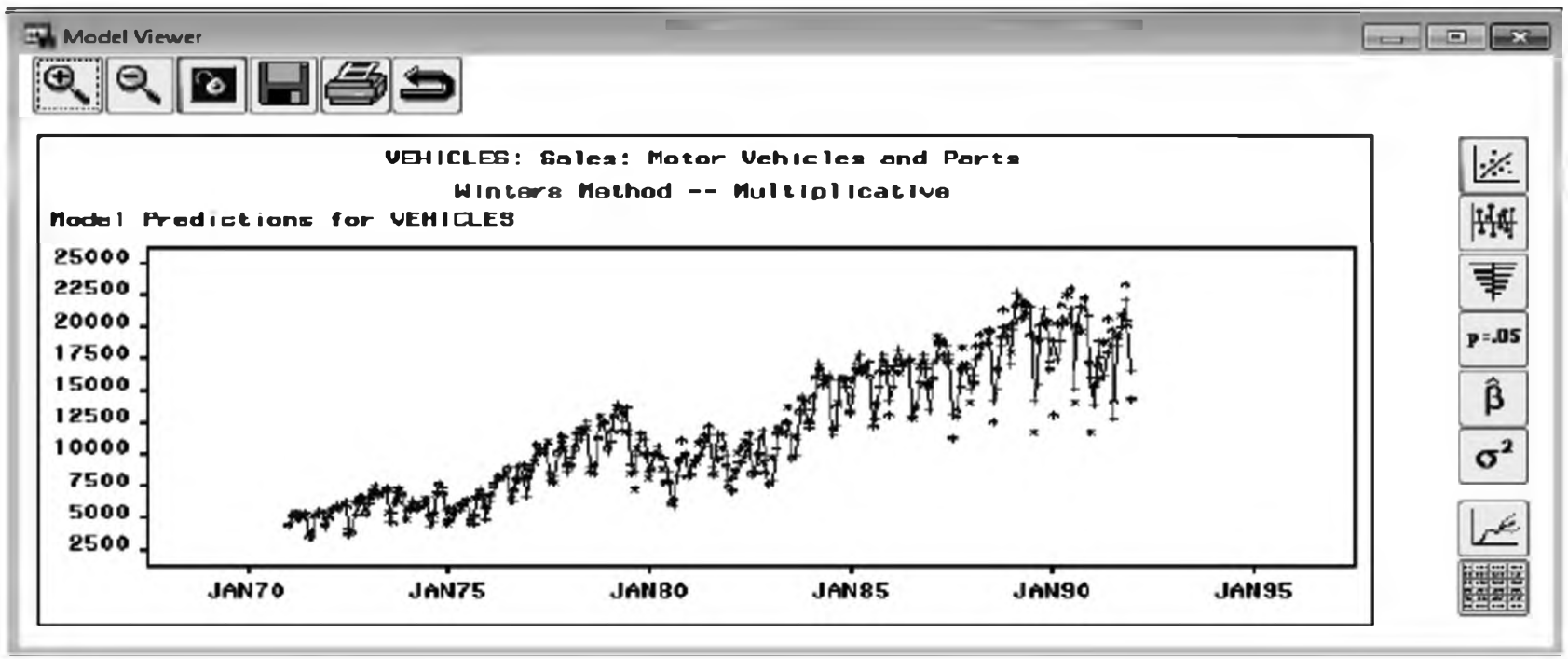


Figura 11-14

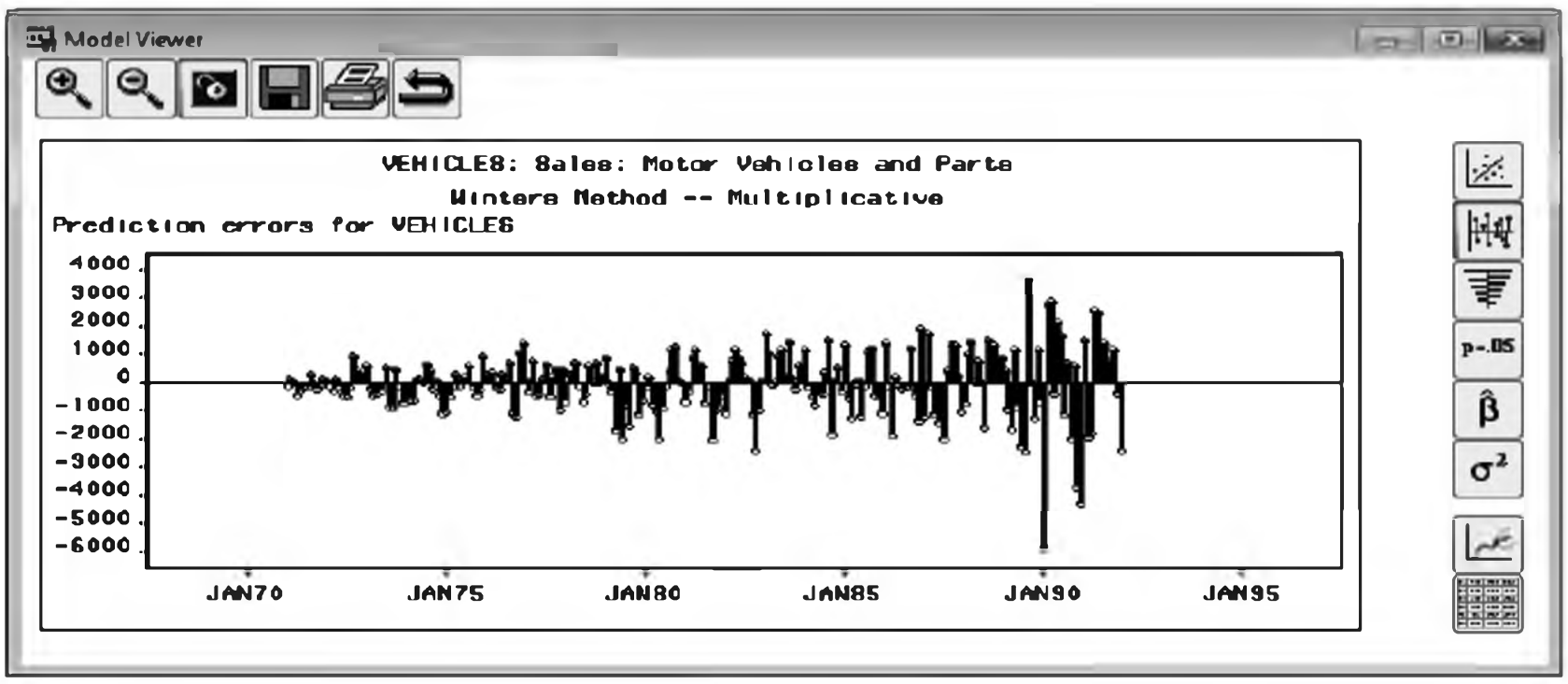


Figura 11-15

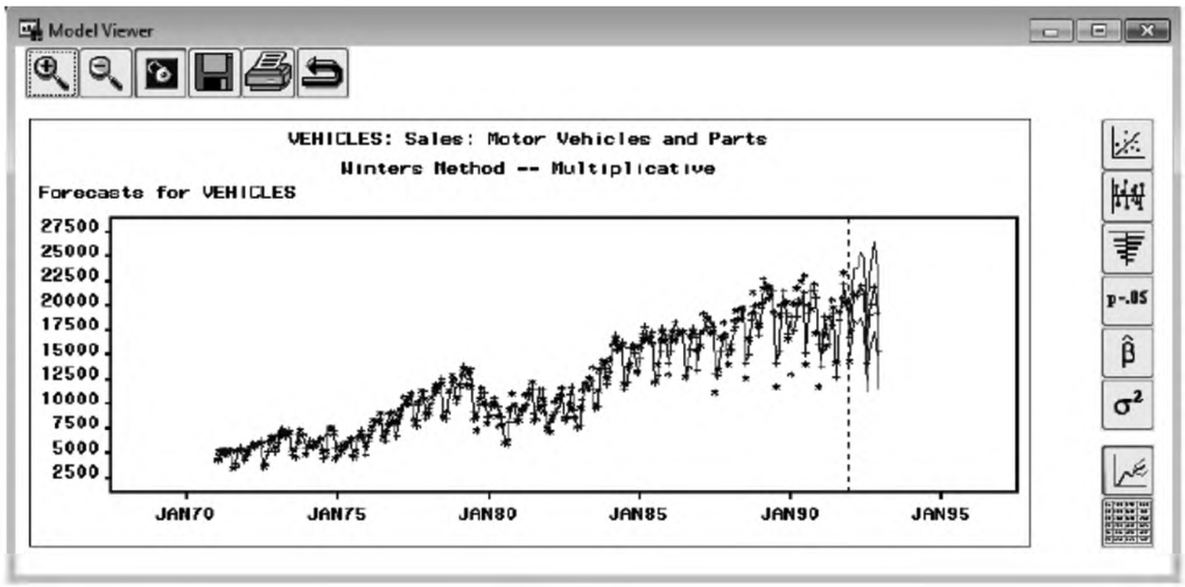
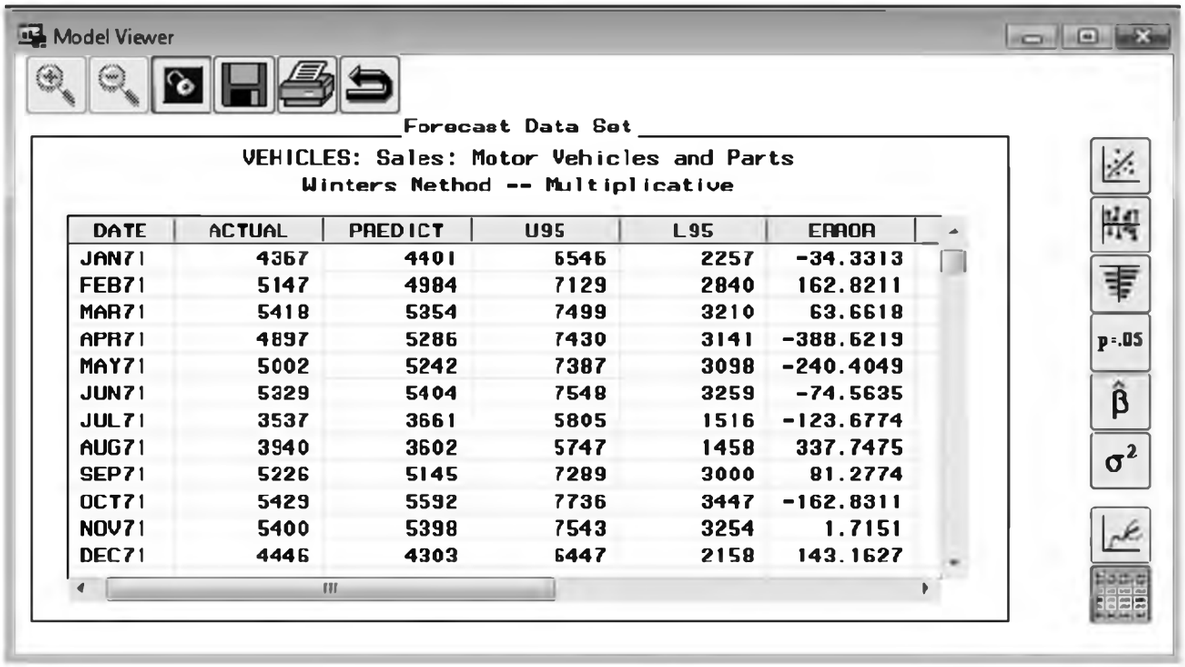


Figura 11-16



DATE	ACTUAL	PREDICT	U95	L95	ERROR
JAN71	4367	4401	6546	2257	-34.3313
FEB71	5147	4984	7129	2840	162.8211
MAR71	5418	5354	7499	3210	63.6618
APR71	4897	5286	7430	3141	-388.6219
MAY71	5002	5242	7387	3098	-240.4049
JUN71	5329	5404	7548	3259	-74.5635
JUL71	3537	3661	5805	1516	-123.6774
AUG71	3940	3602	5747	1458	337.7475
SEP71	5226	5145	7289	3000	81.2774
OCT71	5429	5592	7736	3447	-162.8311
NOV71	5400	5398	7543	3254	1.7151
DEC71	4446	4303	6447	2158	143.1627

Figura 11-17

Vamos a desarrollar ahora un ejemplo para una serie estacional. Consideramos el archivo estacional.sas7bdat que contiene la variable X que define los datos de una serie temporal de estacionalidad mensual desde el 1 de enero de 1968 hasta el 1 de octubre de 1981. Se trata de suavizar la serie temporal X y realizar predicciones para un año completo mediante un método determinista utilizando el modelizador de SAS.

©Alfaomega-RC Libros

Se accede al modelizador de SAS mediante *Soluciones* → *Análisis* → *Sistema de predicción de series temporales* (Figura 11-2). Se obtiene la Figura 11-18 en la que hay que elegir el conjunto de datos que contiene la serie temporal a analizar *Data Set*, la variable que contiene la secuencia temporal *Time ID* y el intevalo estacional *Interval*. Ya hemos visto anteriormente en este capítulo que la serie es estacional mensual. Para ello se hace clic en *Browse* a la derecha de *Data Set* y se elige librería, conjunto de datos, variable Id e intervalo estacional (Figura 11-19). Se hace clic en OK y ya se memorizan las características elegidas (Figura 11-20). Se hace clic en *Produce Forecast* y a la derecha del campo *Series to Process* de la Figura 11-21 se hace clic en *Select* para elegir la variable para la que hay que realizar las predicciones de entre la lista de todas las variables del fichero.

A continuación se hace clic con el botón derecho del ratón sobre cualquier zona de la pantalla y en el menú emergente resultante se elige *Options*. Se despliega el menú de la Figura 11-22 cuyas opciones permitirán administrar el modelizador.

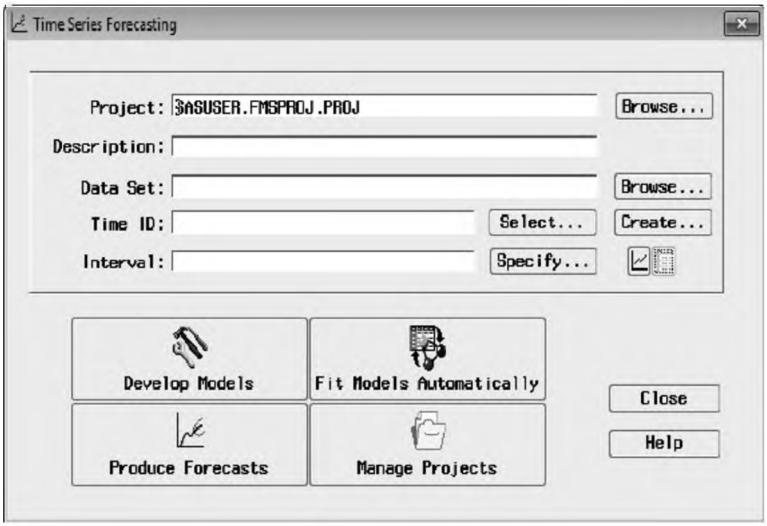


Figura 11-18

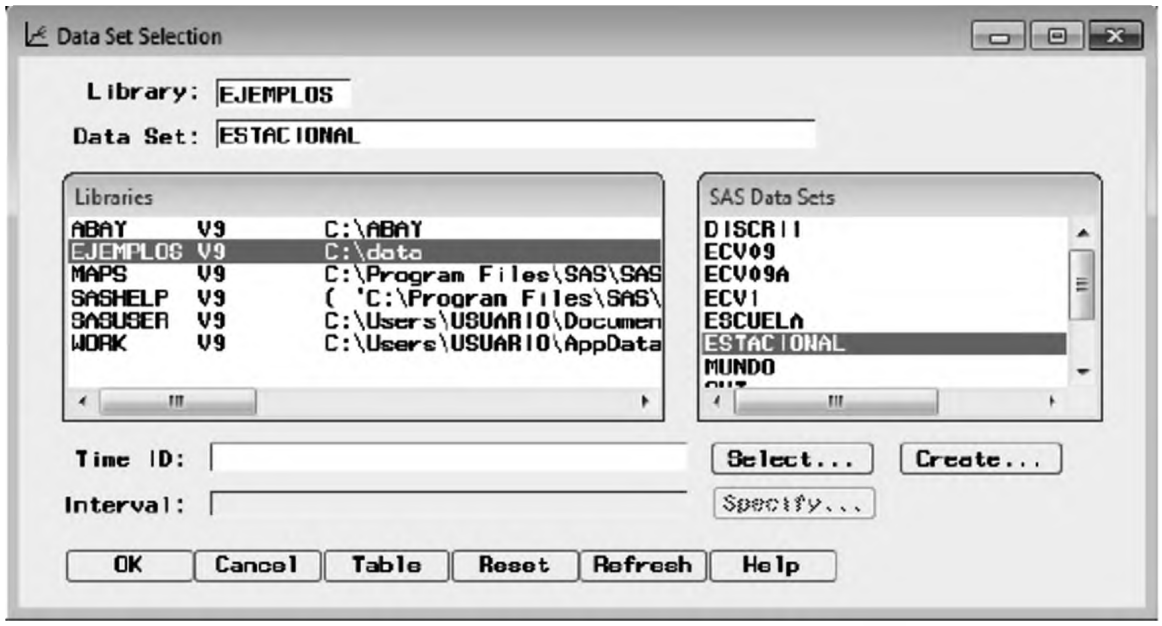


Figura 11-19

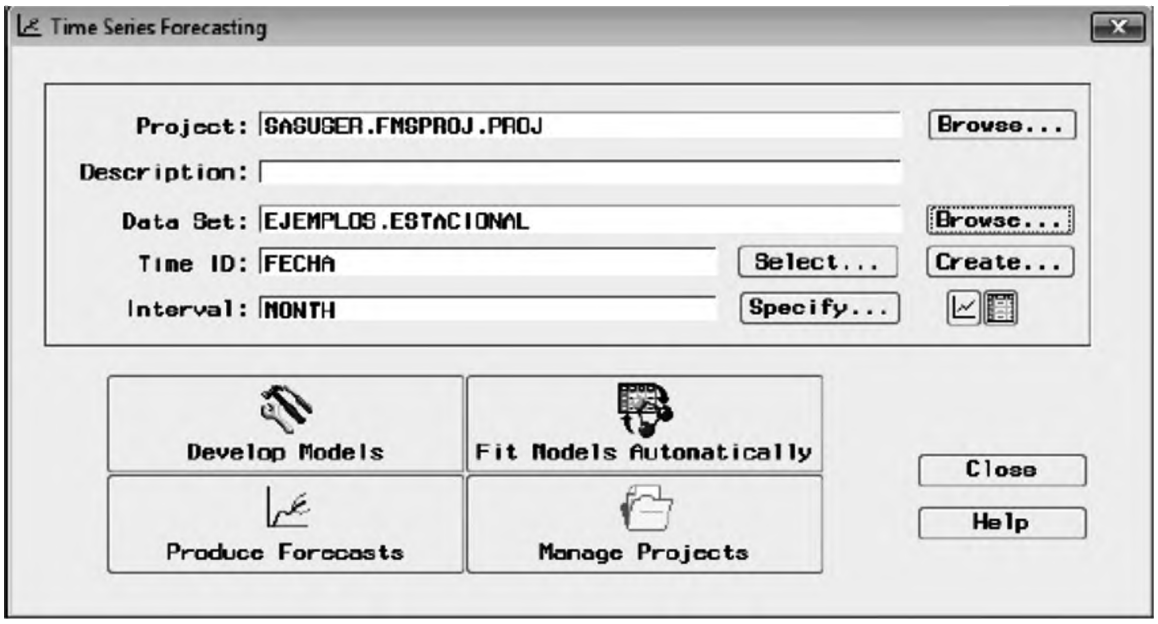


Figura 11-20

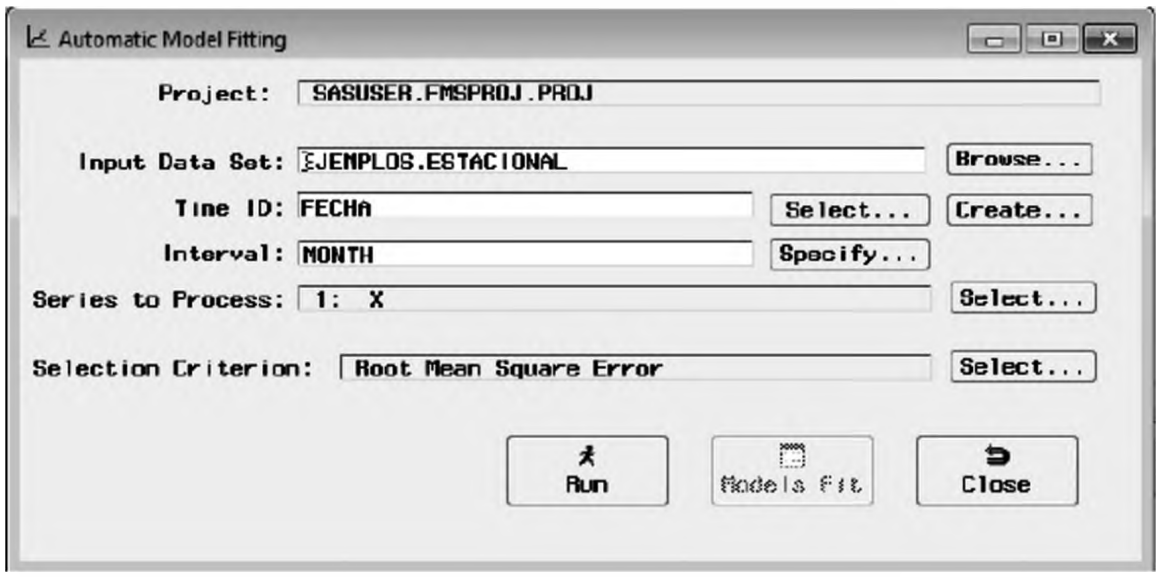


Figura 11-21

©Alfaomega-RC Libros

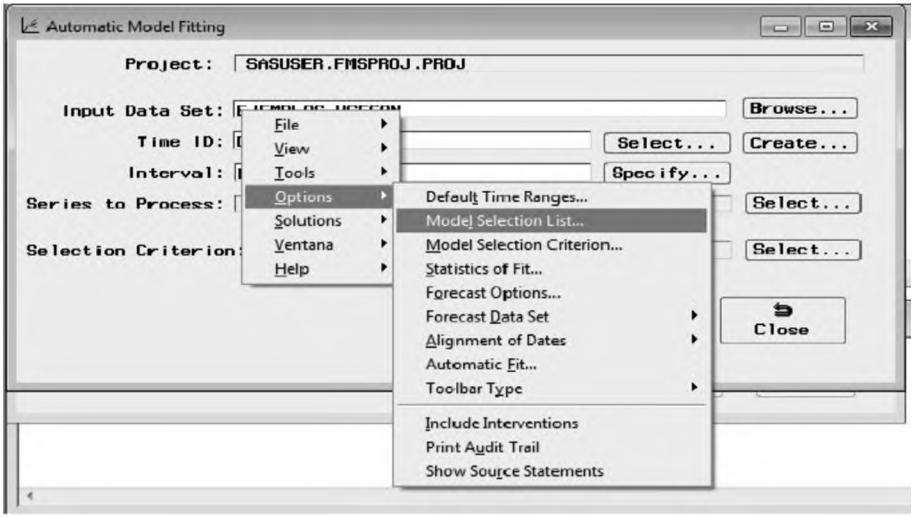


Figura 11-22

La opción *Default Time Ranges* de la Figura 11-22 permite fijar el horizonte de predicción y definir el período de ajuste (Figura 11-23). La opción *Model Selection List* permite elegir los tipos de modelos para suavizar la serie y realizar las predicciones (Figura 11-24). En nuestro caso elegimos los posibles modelos incondicionales deterministas para que el modelizador elija el mejor. La opción *Statistics Fit selection* permite elegir los estadísticos de ajuste para la elección del mejor modelo de entre los propuestos (Figura 11-25). La opción *Model Selection Criterion* presenta los criterios seleccionados para la elección del mejor modelo de ajuste (Figura 11-26). La opción *Forecast Options* permite elegir los límites de confianza para las predicciones (Figura 11-27).

Elegidas todas estas características, se hace clic en *Run* y se obtiene que el mejor modelo es el suavizado exponencial estacional para la serie en logaritmos (Figura 11-28). Seleccionado este modelo, se hace clic en *Graph* y se obtiene la gráfica de la serie (Figura 11-29).

Los iconos de la franja vertical de la derecha de esta figura permiten ver la gráfica de la serie (Figura 11-29), ver la gráfica de los residuos (Figura 11-30), ver la gráfica de la serie original, la serie suavizada y las predicciones en el mismo gráfico (Figura 11-31) y ver una tabla con los valores originales de la serie, los valores predichos, los errores y los intervalos de confianza para las predicciones (Figura 11-32)

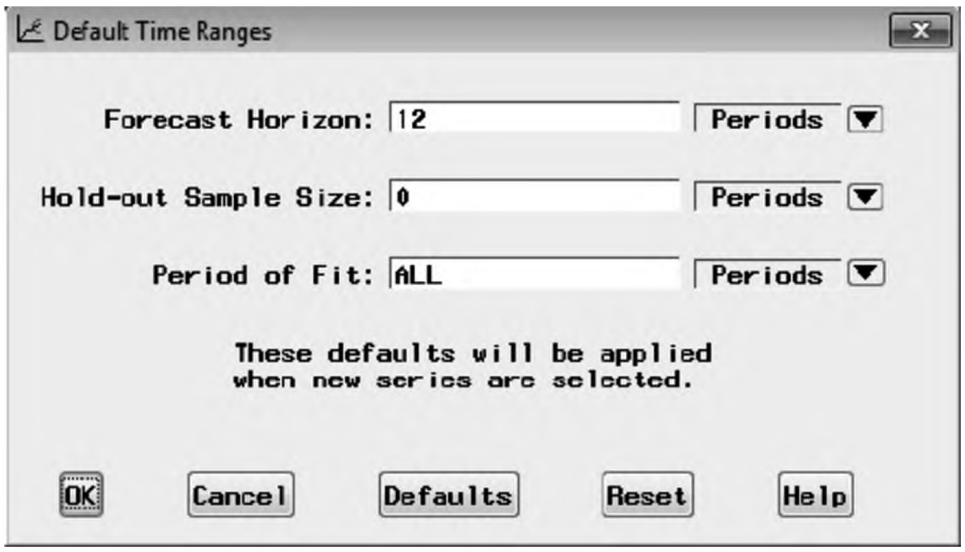


Figura 11-23

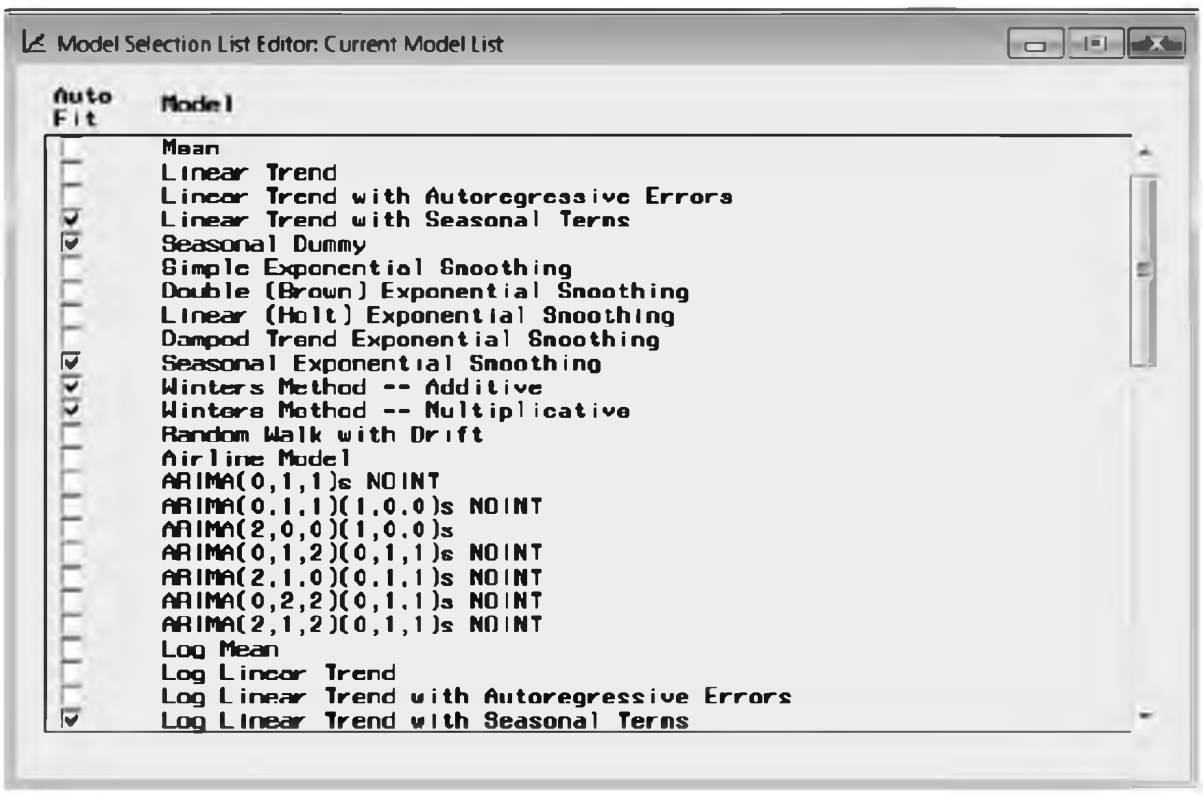


Figura 11-24

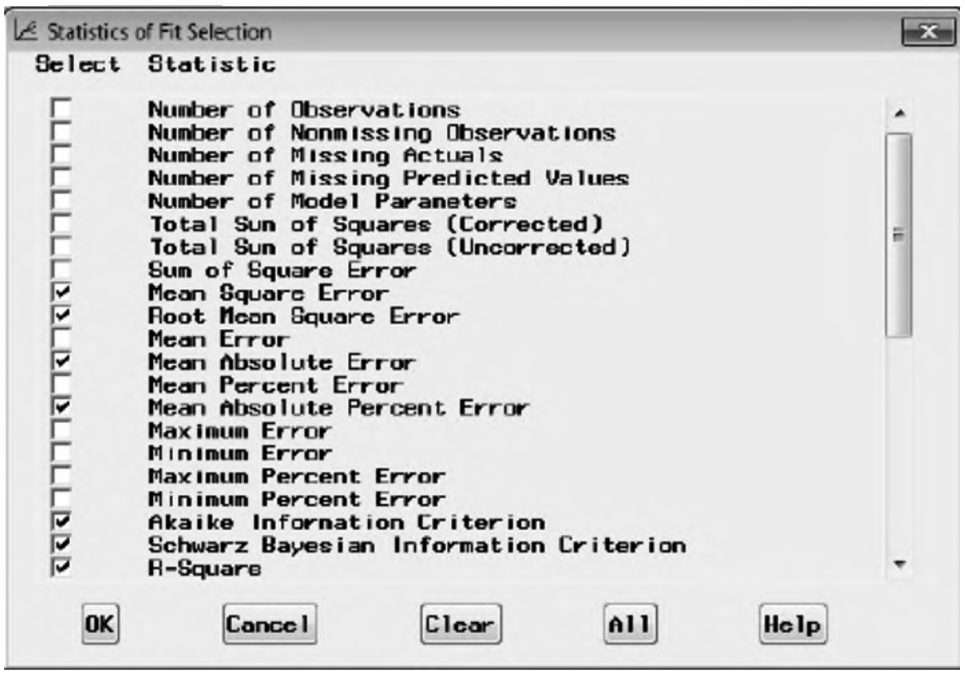


Figura 11-25

©Alfaomega-RC Libros

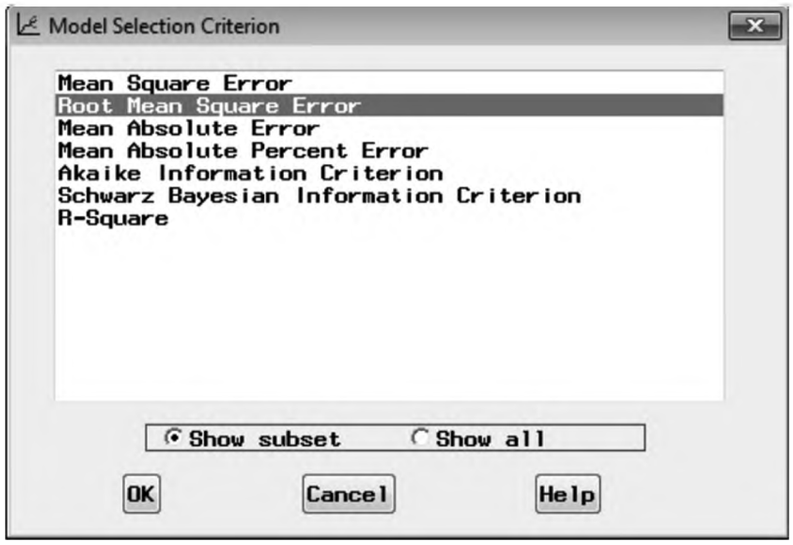


Figura 11-26

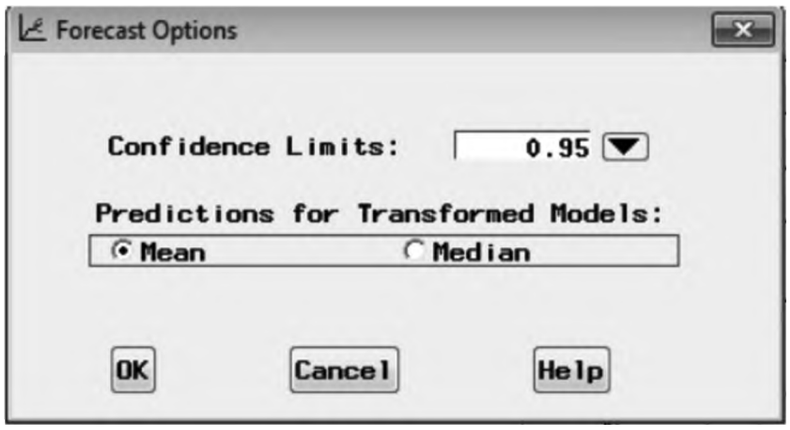


Figura 11-27

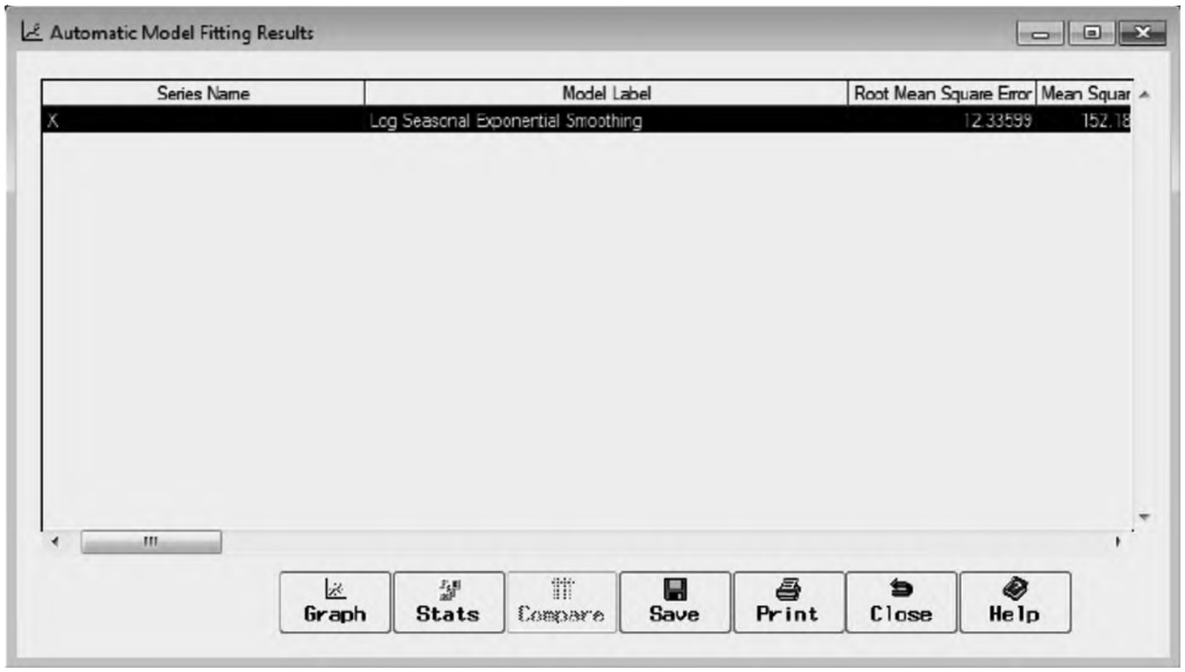


Figura 11-28

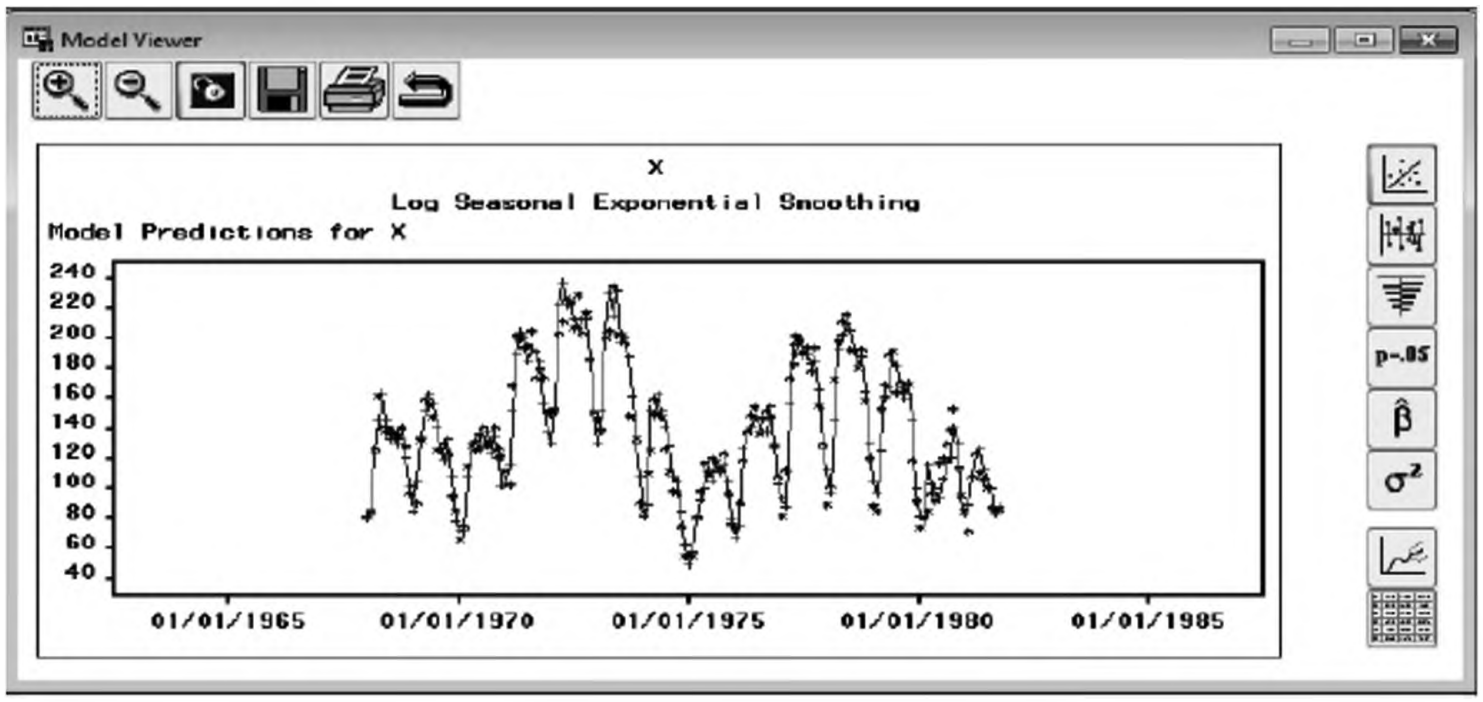


Figura 11-29

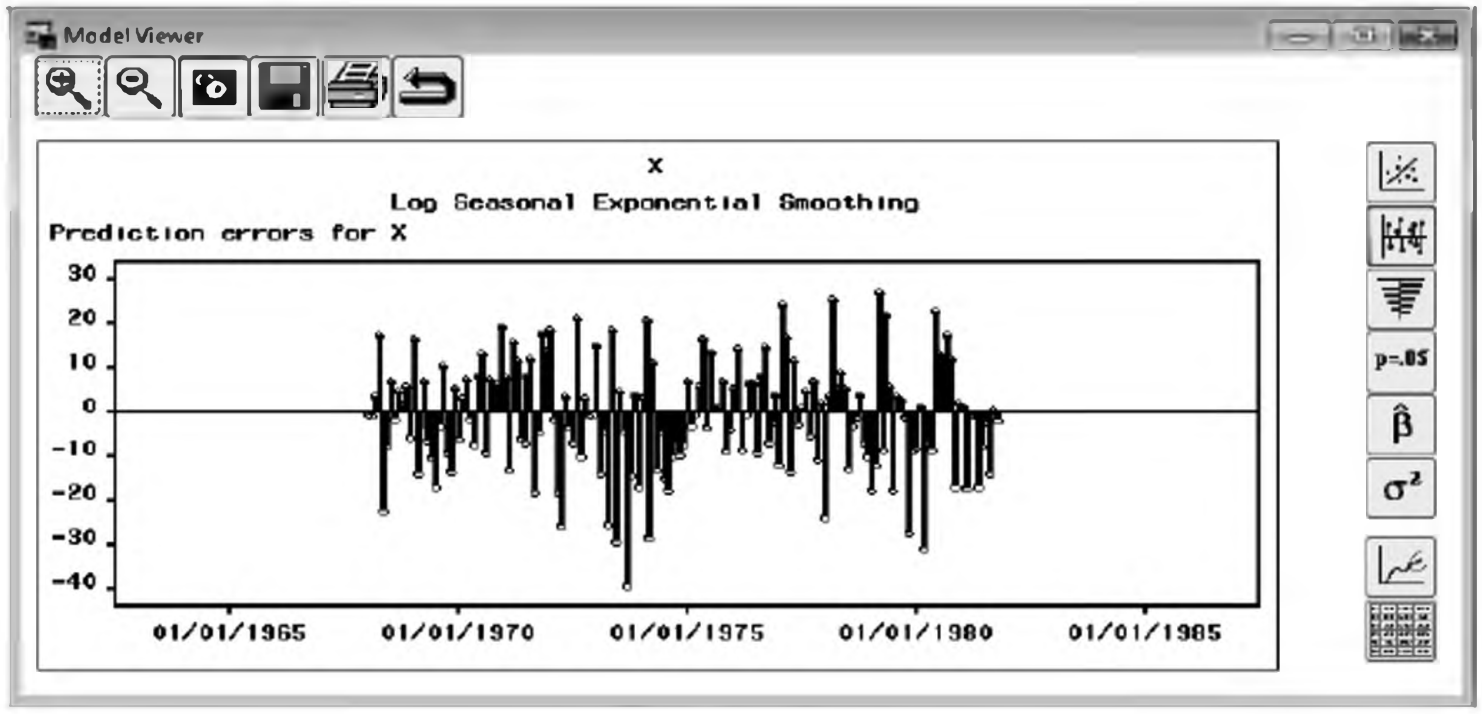


Figura 11-30

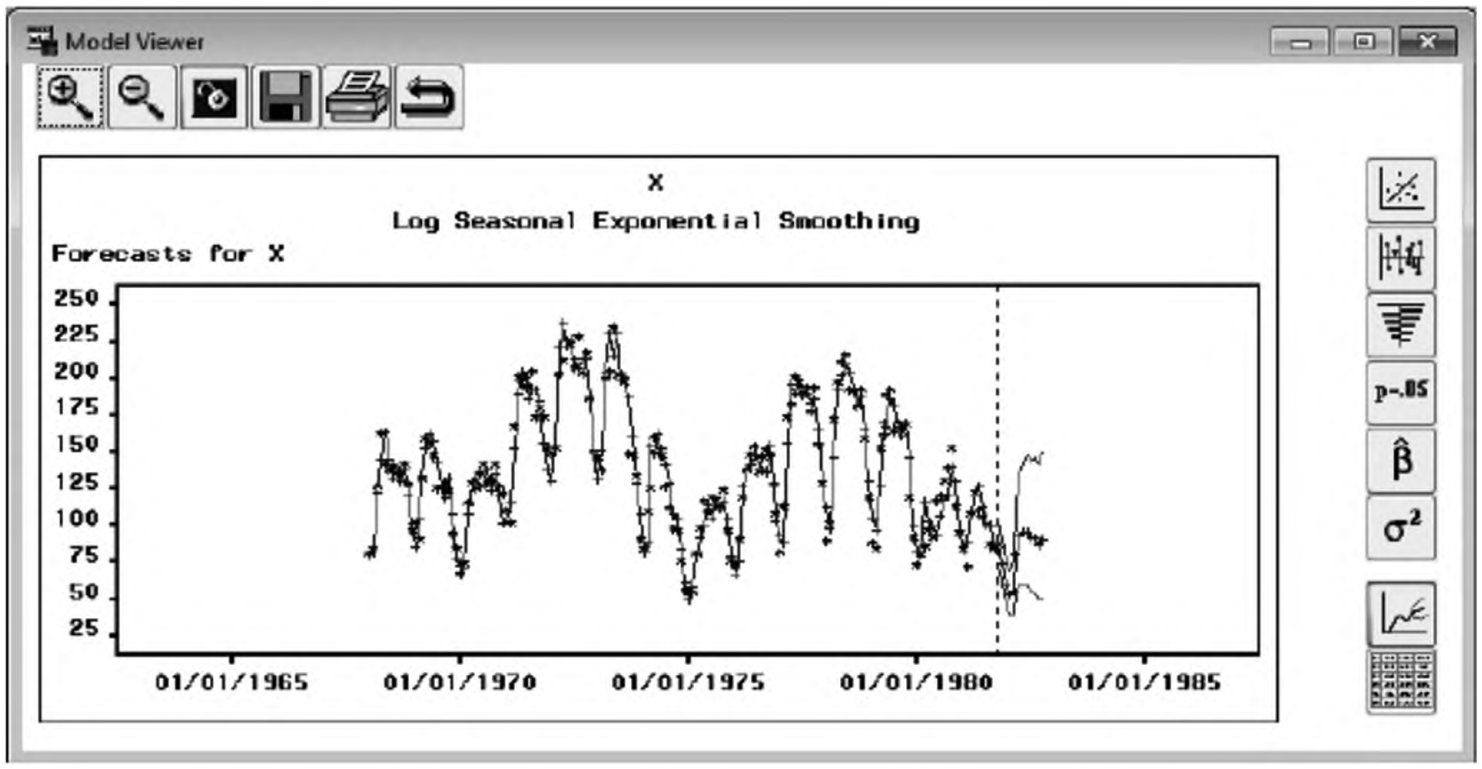
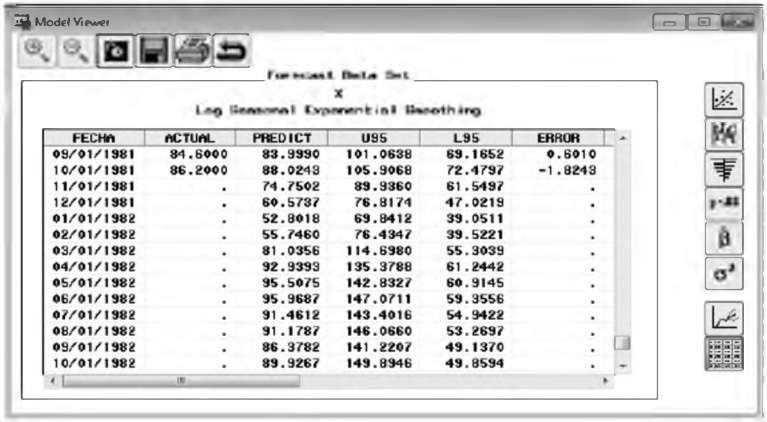


Figura 11-31

©Alfaomega-RC Libros



FECHA	ACTUAL	PREDICT	U95	L95	ERROR
09/01/1981	84.6000	83.9990	101.0638	69.1652	0.6010
10/01/1981	86.2000	88.0243	105.9068	72.4797	-1.8243
11/01/1981	.	74.7502	89.9360	61.5497	.
12/01/1981	.	60.5737	76.8174	47.0219	.
01/01/1982	.	52.8018	69.8412	39.0511	.
02/01/1982	.	55.7460	76.4347	39.5221	.
03/01/1982	.	81.0356	114.6980	55.3039	.
04/01/1982	.	92.9393	135.3788	61.2442	.
05/01/1982	.	95.5075	142.8327	60.9145	.
06/01/1982	.	95.9687	147.0711	59.3556	.
07/01/1982	.	91.4612	143.4016	54.9422	.
08/01/1982	.	91.1787	146.0660	53.2697	.
09/01/1982	.	86.3782	141.2207	49.1370	.
10/01/1982	.	89.9267	149.8946	49.8594	.

Figura 11-32

Comparando las predicciones obtenidas en la Figura 11-32 con las obtenidas en la Figura 11-16 se observa que no hay muchas diferencias. Los distintos modelizadores no tienen por qué seleccionar los mismos modelos de ajuste. Sin embargo, las predicciones difieren en muy poco.

MODELOS ARIMA CON EL MODELIZADOR AUTOMÁTICO DE SAS

Ya sabemos que SAS dispone de un modelizador autmático que permite ajustar series a modelos ARIMA, suavizarlas y obtener predicciones. Se accede al modelizador mediante *Soluciones → Análisis → Sistema de predicción de series temporales* (Figura 11-2). Se obtiene la Figura 11-33 en la que hay que elegir el conjunto de datos que contiene la serie temporal a analizar *Data Set*, la variable que contiene la secuencia temporal *Time ID* y el intervalo estacional *Interval*. Para ello se hace clic en *Browse* a la derecha de *Data Set* y se elige librería, conjunto de datos, variable Id e intervalo estacional (Figura 11-34). Se hace clic en OK y ya se memorizan las características elegidas. Como ya sabemos que nuestra serie no es estacional, SAS permite crear una varible ID de nombre DATE y de intervalo 1 haciendo clic en el botón *Create* situado a la derecha del campo *Time ID* y eligiendo *Create from observation numbers* (Figura 11-35). Se hace clic en *Fit Models Automaticaly* y a la derecha del campo *Series to Process* de la Figura 11-36 se hace clic en *Select* para elegir la variable para la que hay que realizar las predicciones de entre la lista de todas las variables del fichero.

A continuación se hace clic con el botón derecho del ratón sobre cualquier zona de la pantalla y en el menú emergente resultante se elige *Options*. Se despliega el menú de la Figura 11-37 cuyas opciones permitirán administrar el modelizador.

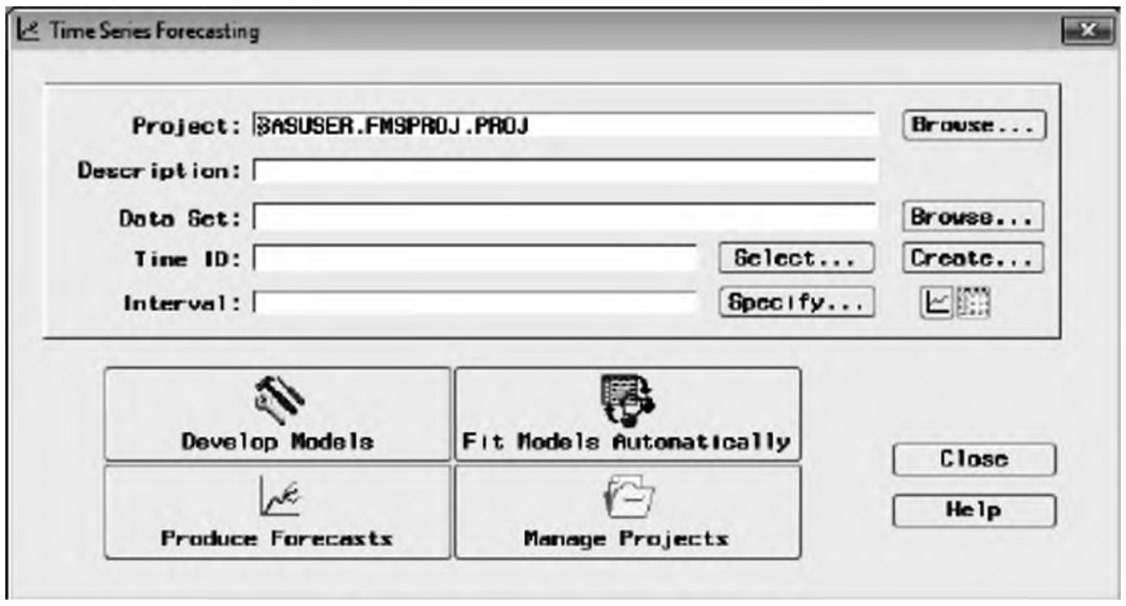


Figura 11-33

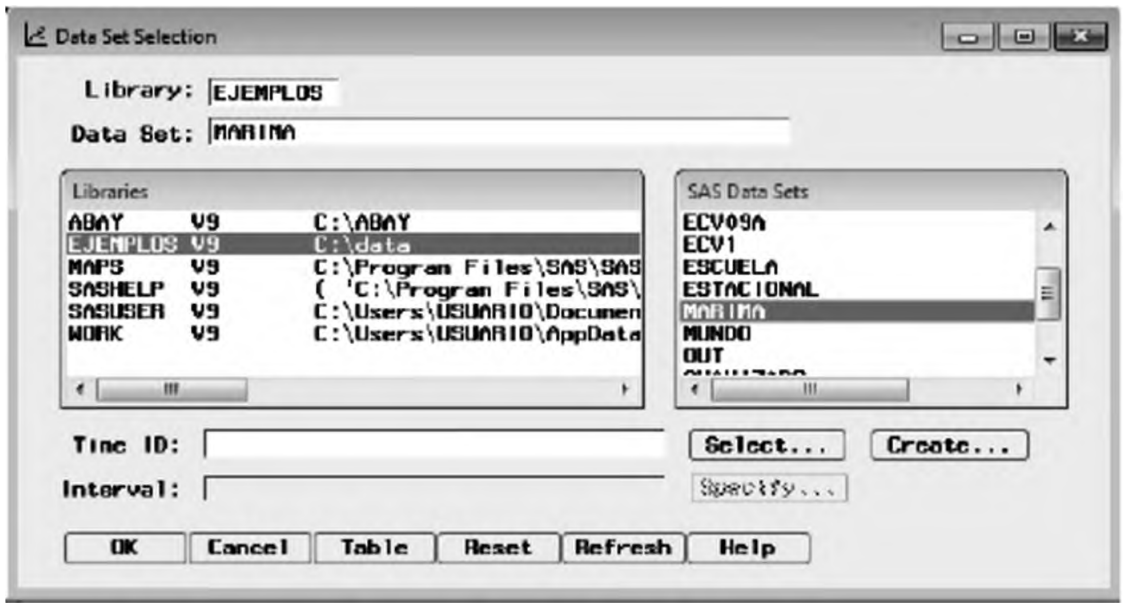


Figura 11-34

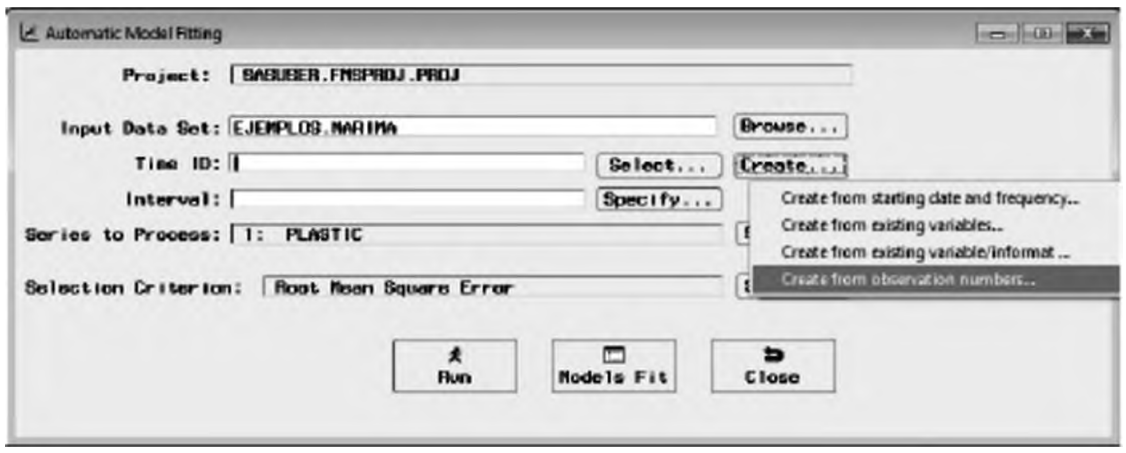


Figura 11-35

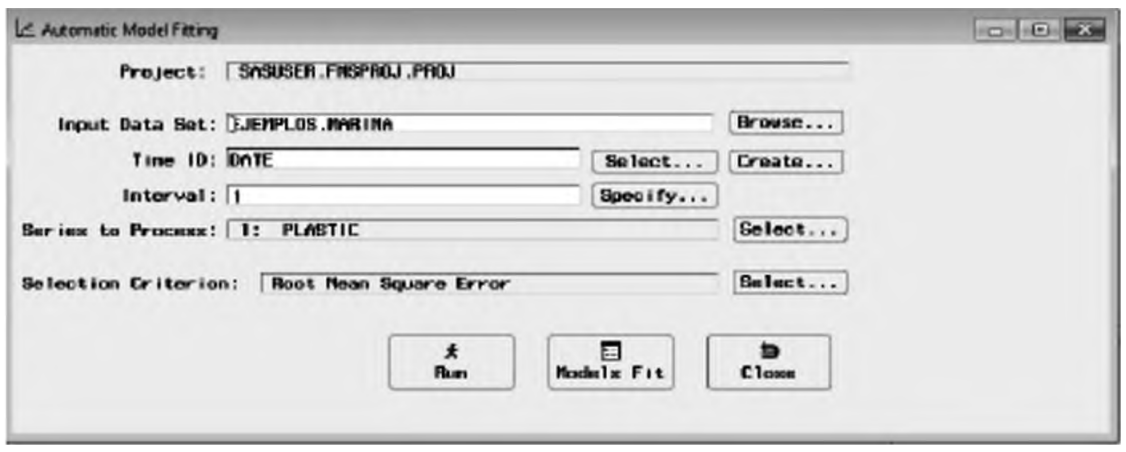


Figura 11-36

©Alfaomega-RC Libros

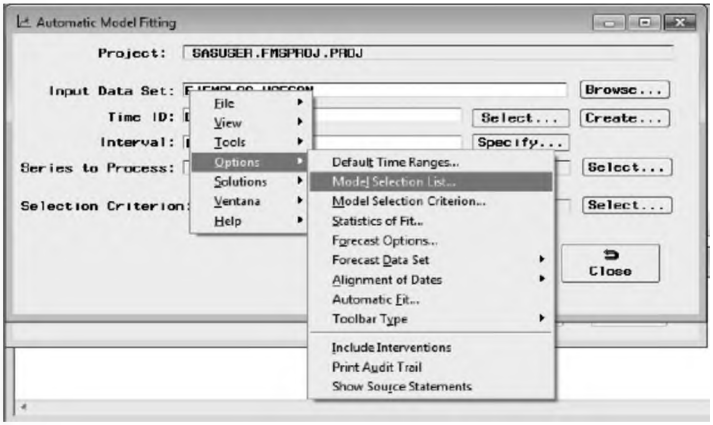


Figura 11-37

La opción *Default Time Ranges* de la Figura 11-37 permite fijar el horizonte de predicción y definir el período de ajuste (Figura 11-38). La opción *Model Selection List* permite elegir los tipos de modelos para suavizar la serie y realizar las predicciones (Figura 11-39). En nuestro caso elegimos todos los posibles modelos para que el modelizador elija el mejor. La opción *Statistics Fit selection* permite elegir los estadísticos de ajuste para la elección del mejor modelo de entre los propuestos (Figura 11-40). La opción *Model Selection Criterion* presenta los criterios seleccionados para la elección del mejor modelo de ajuste (Figura 11-41). La opción *Forecast Options* permite elegir los límites de confianza para las predicciones (Figura 11-42). La opción *Automatic Fit* (Figura 11-43) permite definir entre qué conjunto de modelos se hace la selección (todos los seleccionados o los adecuados según la fase de diagnóstico). También permite elegir el número de modelos obtenidos como solución óptima ordenados de mejor a peor. En nuestro caso los elegiremos todos marcando *All Models*. Esta es una buena práctica para darle las máximas posibilidades de elección y ordenación de modelos óptimos para el ajuste de nuestra serie.

Elegidas todas estas características, se hace clic en *Run* y se obtiene que el mejor modelo es un suavizado exponencial con tendencia específico (Figura 11-44). Seleccionado este modelo, se hace clic en *Graph* y se obtiene la gráfica de la serie (Figura 11-45).

Los iconos de la franja vertical de la derecha de esta figura permiten ver la gráfica de la serie (Figura 11-45), ver la gráfica de los residuos (Figura 11-46), ver la gráfica de la serie original, la serie suavizada y las predicciones en el mismo gráfico (Figura 11-47) y ver una tabla con laos valores originales de la serie, los valores predichos, los errores y los intervalos de confianza para las predicciones (Figura 11-48).

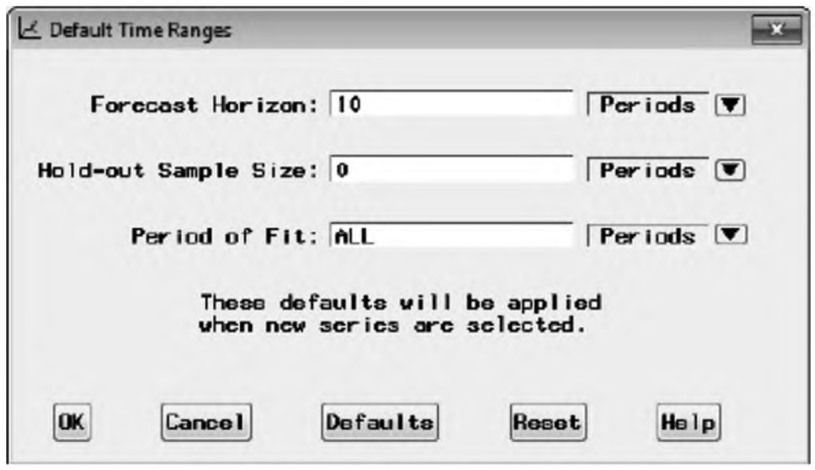


Figura 11-38

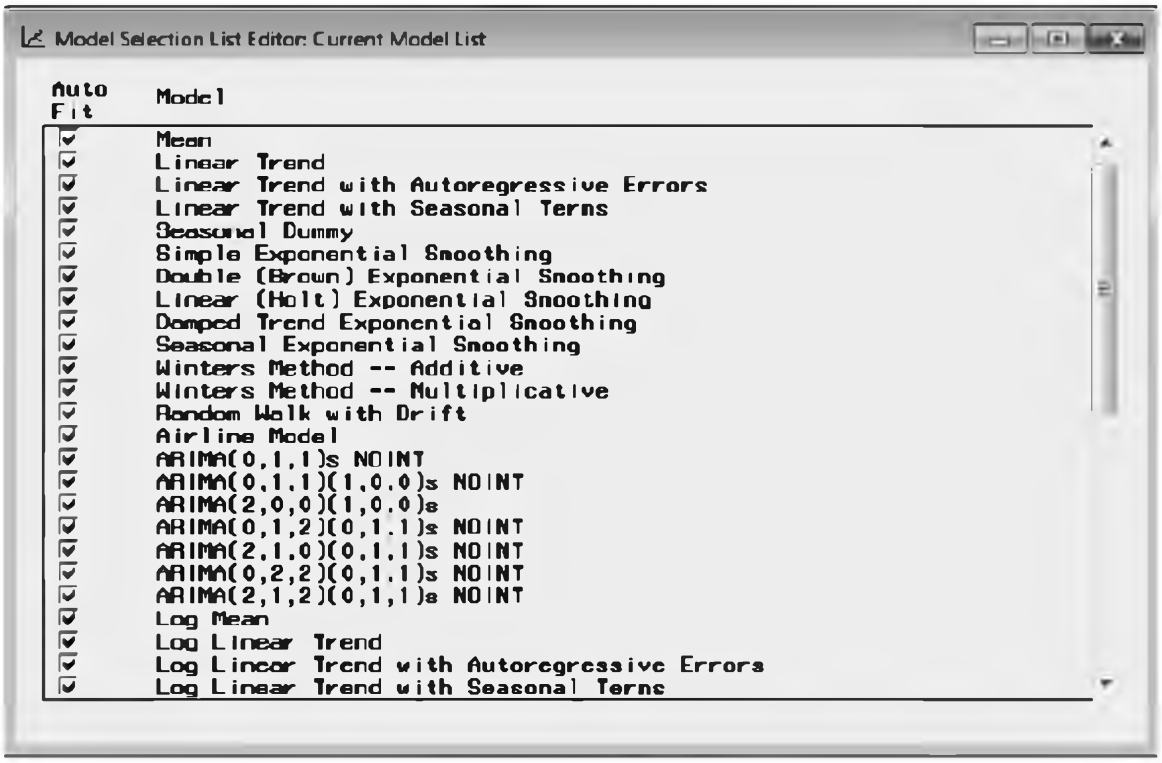


Figura 11-39

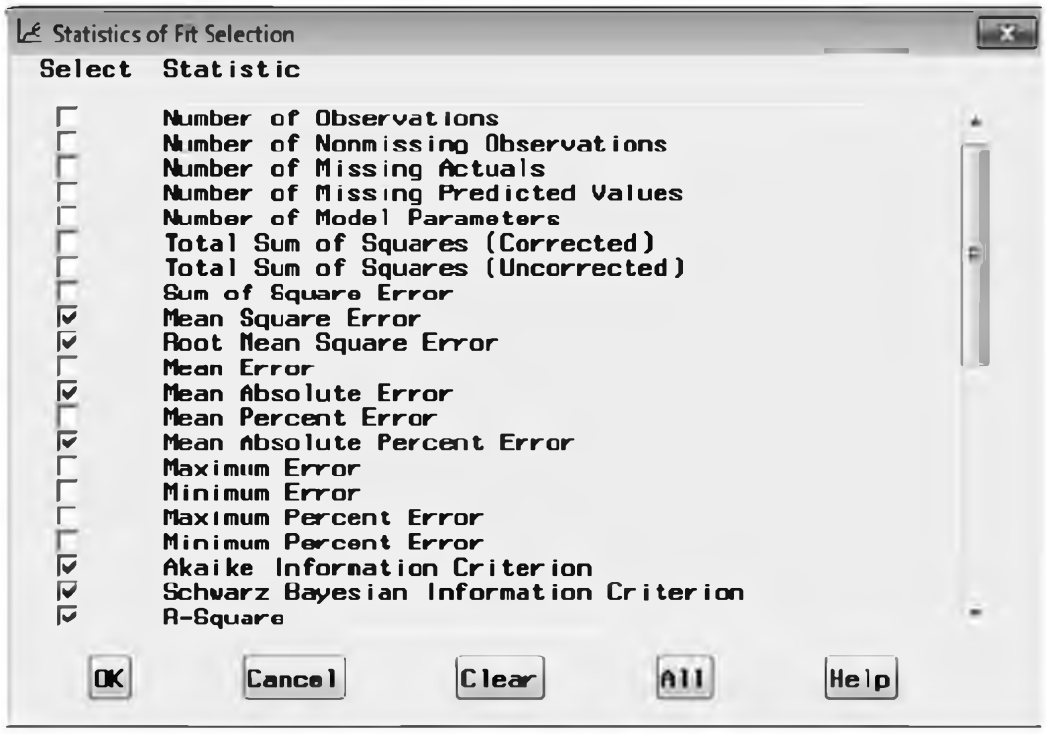


Figura 11-40

©Alfaomega-RC Libros

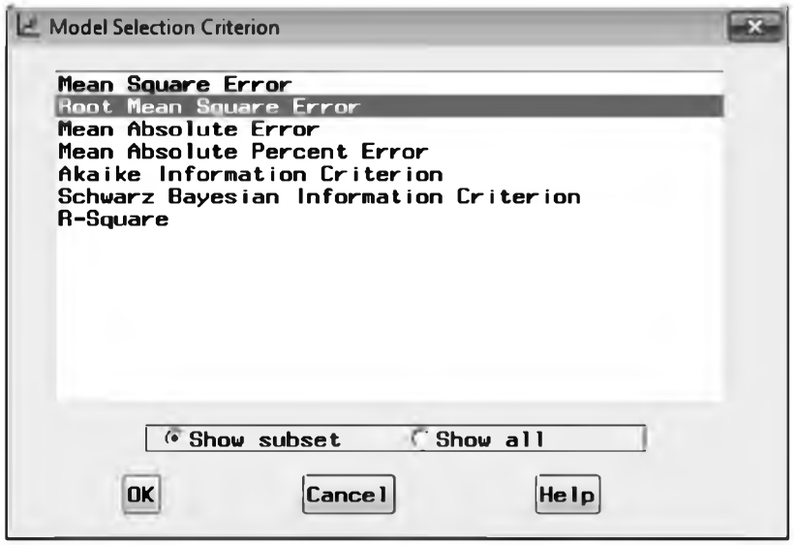


Figura 11-41

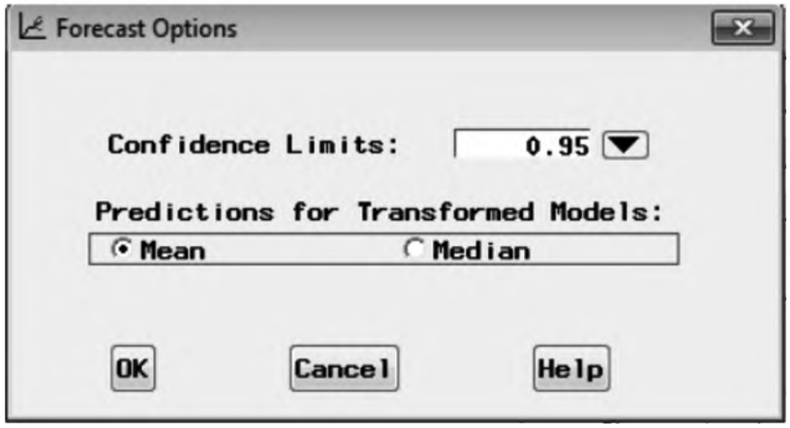


Figura 11-42

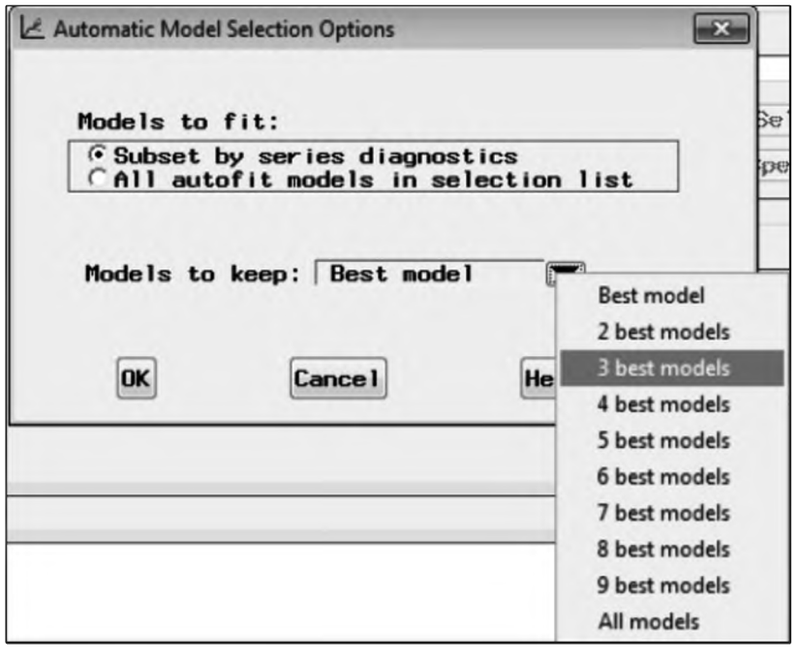
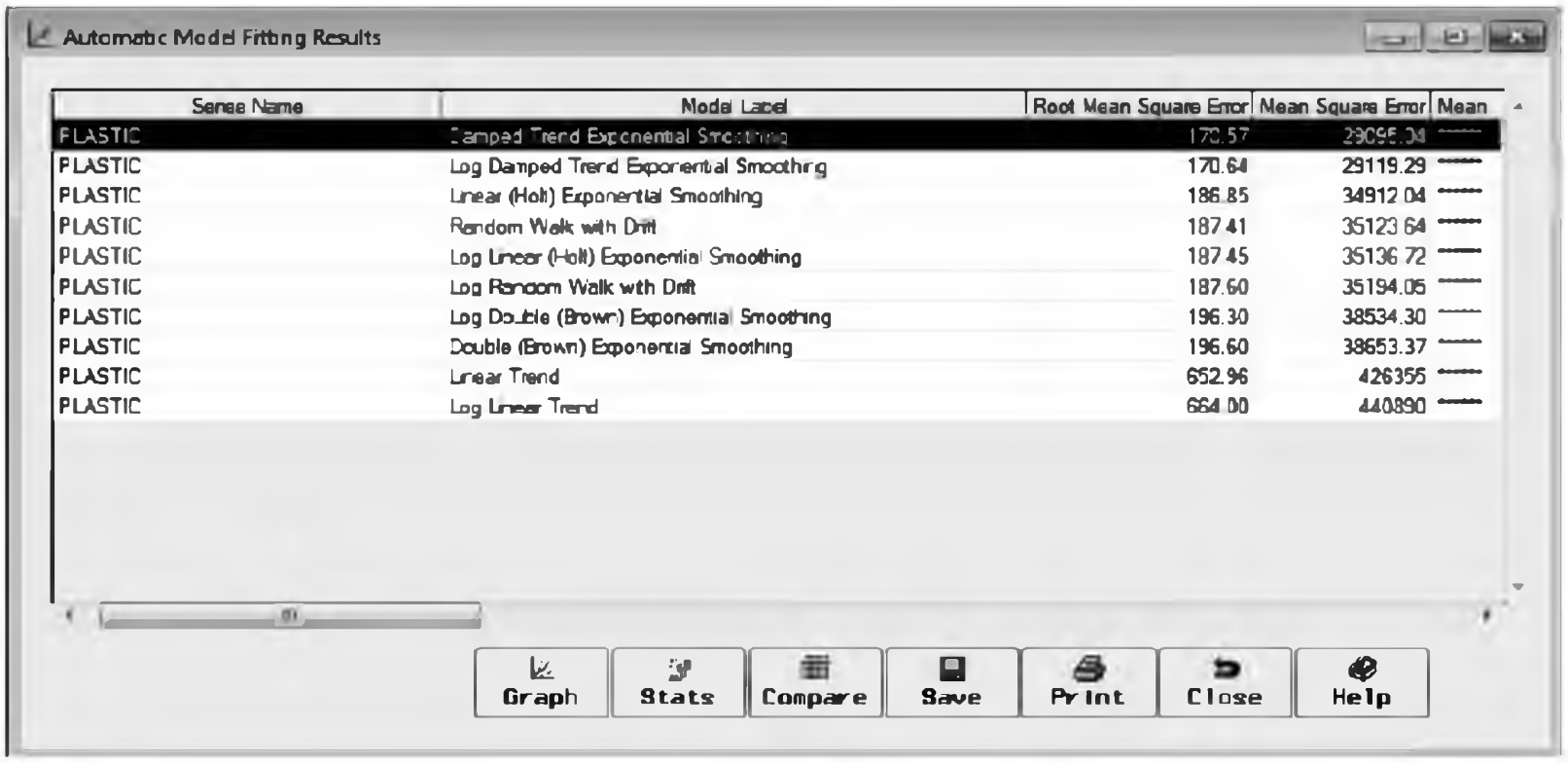


Figura 11-43



Series Name	Model Label	Root Mean Square Error	Mean Square Error	Mean
PLASTIC	Damped Trend Exponential Smoothing	170.57	29095.04	------
PLASTIC	Log Damped Trend Exponential Smoothing	170.64	29119.29	------
PLASTIC	Linear (Holt) Exponential Smoothing	186.85	34912.04	------
PLASTIC	Random Walk with Drift	187.41	35123.64	------
PLASTIC	Log Linear (Holt) Exponential Smoothing	187.45	35136.72	------
PLASTIC	Log Random Walk with Drift	187.60	35194.05	------
PLASTIC	Log Double (Brown) Exponential Smoothing	196.30	38534.30	------
PLASTIC	Double (Brown) Exponential Smoothing	196.60	38653.37	------
PLASTIC	Linear Trend	652.96	426355	------
PLASTIC	Log Linear Trend	664.00	440890	------

Figura 11-44

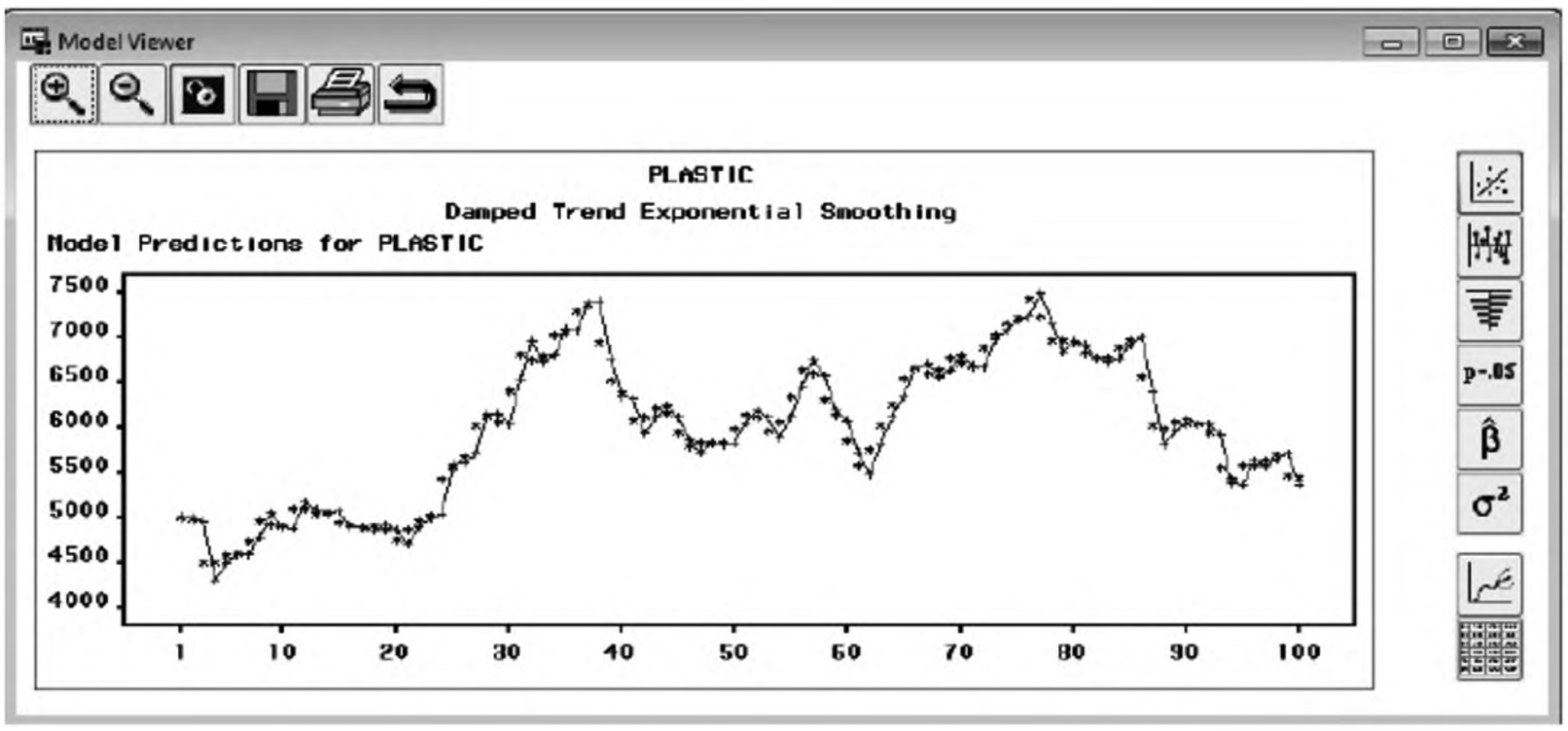


Figura 11-45

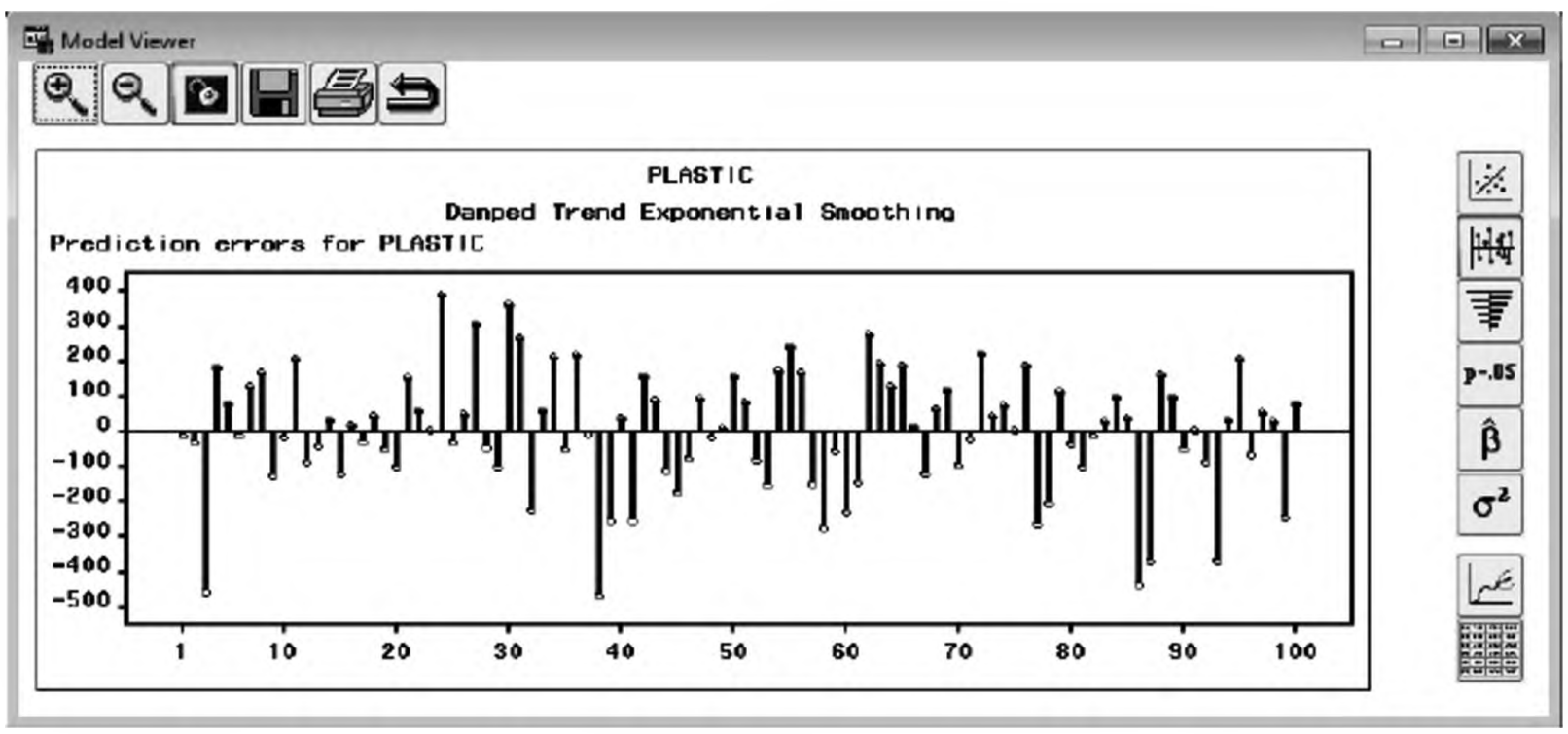


Figura 11-46

©Alfaomega-RC Libros

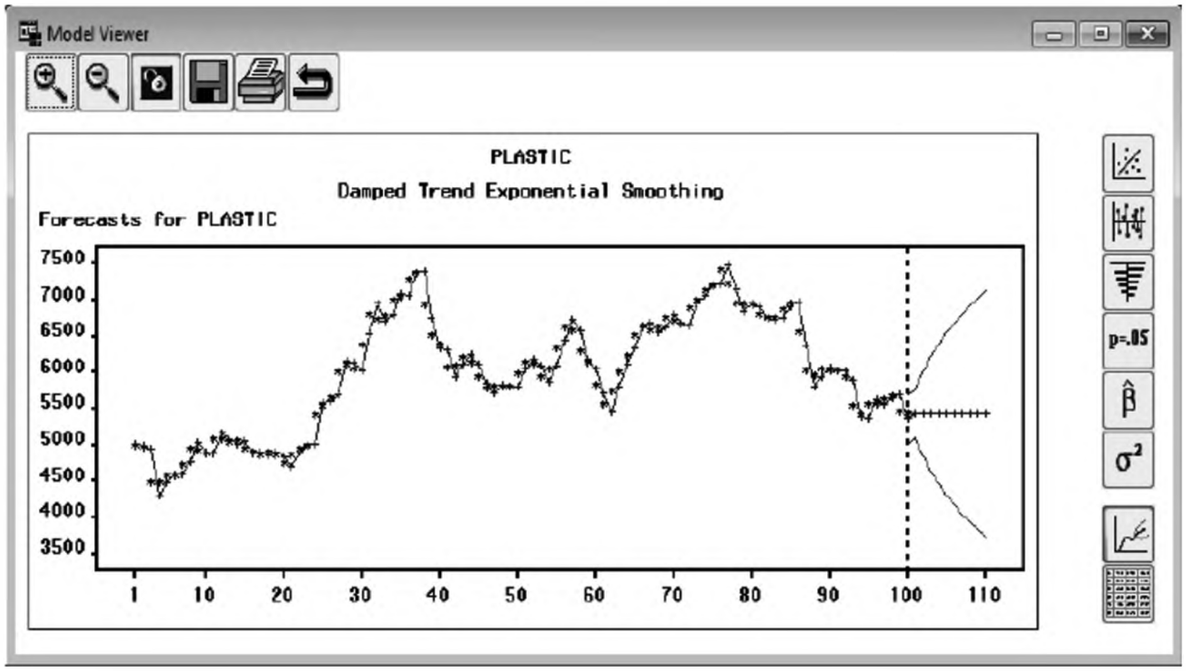
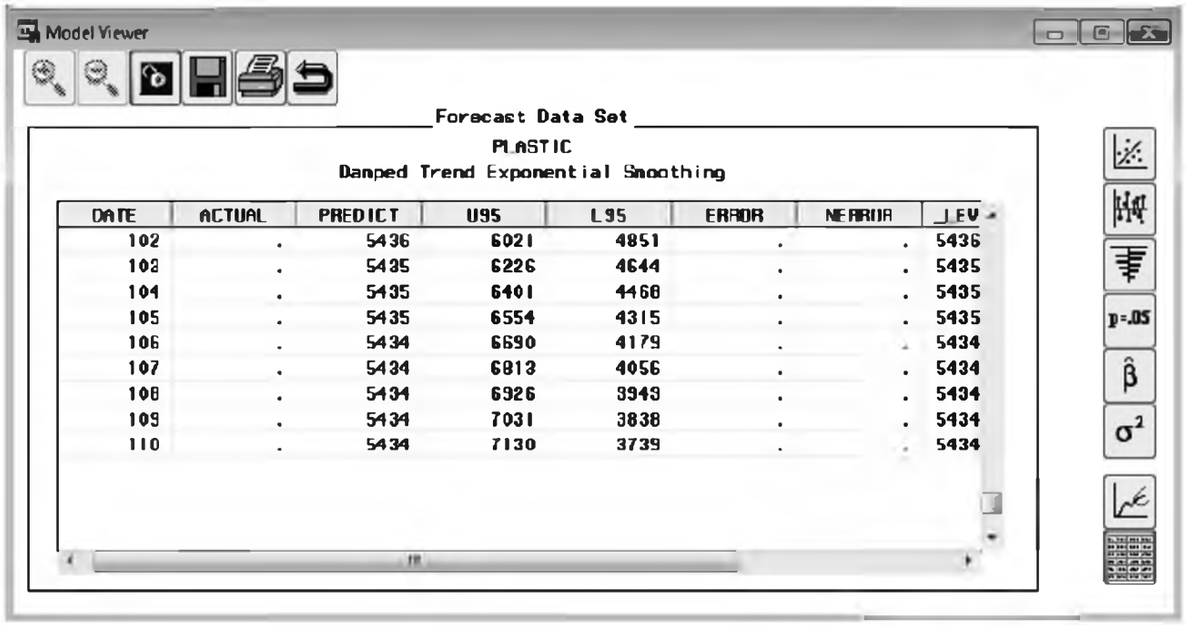


Figura 11-47



DATE	ACTUAL	PREDICT	U95	L95	ERROR	NERROR	LEV
102	.	5436	6021	4851	.	.	5436
103	.	5435	6226	4644	.	.	5435
104	.	5435	6401	4468	.	.	5435
105	.	5435	6554	4315	.	.	5435
106	.	5434	6690	4179	.	.	5434
107	.	5434	6813	4056	.	.	5434
108	.	5434	6926	3943	.	.	5434
109	.	5434	7031	3838	.	.	5434
110	.	5434	7130	3739	.	.	5434

Figura 11-48

Se observa que el modelizador automático de SAS no ha elegido un modelo ARIMA de líneas aéreas, pero observando las predicciones, la solución es muy similar. No todos los predictores automáticos tienen que elegir el mismo modelo para una serie. Hay modelos muy similares.

MODELOS ARIMA ESTACIONALES Y GENERALES CON EL MODELIZADOR AUTOMÁTICO DE SAS

SAS dispone de un modelizador automático que permite ajustar series a modelos ARIMA estacionales y generales, suavizarlas y obtener predicciones. Como ejemplo utilizaremos la variable *x* del fichero *seriesg.sas7bdat* que contiene el número de pasajeros mensuales en un aeropuerto. Se accede al modelizador mediante *Soluciones* → *Análisis* → *Sistema de predicción de series temporales*. Se obtiene la Figura 11-49 en la que hay que elegir el conjunto de datos que contiene la serie temporal a analizar *Data Set*, la variable que contiene la secuencia temporal *Time ID* y el intervalo estacional *Interval*. Para ello se hace clic en *Browse* a la derecha de *Data Set* y se elige librería, conjunto de datos, variable Id e intervalo estacional. Se hace clic en OK y ya se memorizan las características elegidas. Se hace clic en *Fit Models Automaticaly* y a la derecha del campo *Series to Process* de la Figura 11-50 se hace clic en *Select* para elegir la variable para la que hay que realizar las predicciones de entre la lista de todas las variables del fichero (en nuestro caso la variable *x*). A continuación se hace clic con el botón derecho del ratón sobre cualquier zona de la pantalla y en el menú emergente resultante se elige *Options*. Se despliega el menú de la Figura 11-51 cuyas opciones permitirán administrar el modelizador.

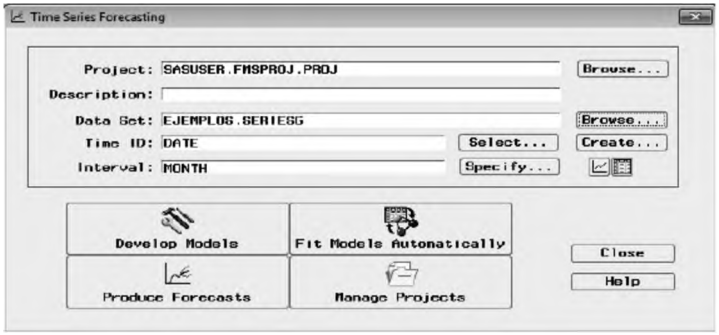


Figura 11-49

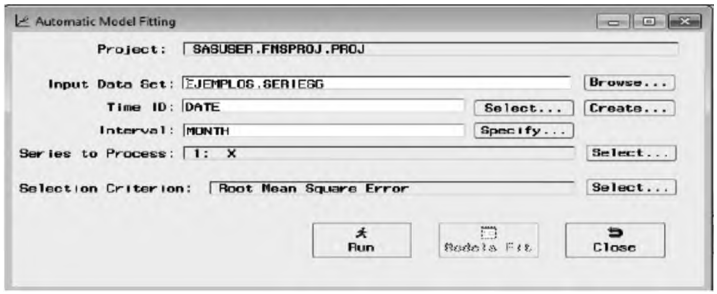


Figura 11-50

©Alfaomega-RC Libros

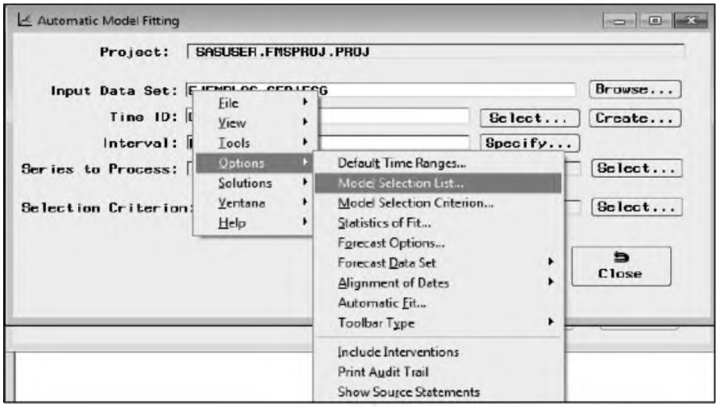


Figura 11-51

La opción *Default Time Ranges* de la Figura 11-51 permite fijar el horizonte de predicción y definir el período de ajuste. La opción *Model Selection List* permite elegir los tipos de modelos para suavizar la serie y realizar las predicciones (Figura 11-52). En nuestro caso elegimos todos los posibles modelos ARIMA para que el modelizador escoja el mejor. La opción *Statistics Fit selection* permite elegir los estadísticos de ajuste para la elección del mejor modelo de entre los propuestos. La opción *Model Selection Criterion* presenta los criterios seleccionados para la elección del mejor modelo de ajuste. La opción *Forecast Options* permite elegir los límites de confianza para las predicciones. La opción *Automatic Fit* (Figura 11-53) permite definir entre qué conjunto de modelos se hace la selección (todos los seleccionados o los adecuados según la fase de diagnóstico). También permite elegir el número de modelos obtenidos como solución óptima ordenados de mejor a peor. En nuestro caso elegiremos los tres mejores modelos. Esta es una buena práctica para darle las máximas posibilidades de elección y ordenación de modelos óptimos para el ajuste de nuestra serie.

Elegidas todas estas características, se hace clic en *Run* y se obtienen los tres mejores modelos (Figura 11-54). De entre los tres candidatos resulta muy adecuado el modelo de líneas aéreas en logaritmos (es el que mejor significatividad de parámetros estimados tiene). Seleccionado este modelo, se hace clic en *Graph* y se obtiene la gráfica de la serie. Los iconos de la franja vertical de la derecha de esta figura permiten ver la gráfica de la serie, la gráfica de los residuos (Figura 11-56), Los parámetros estimados (Figura 11-57), la gráfica de la serie original, la serie suavizada y las predicciones en el mismo gráfico (Figura 11-55) y una tabla con los valores originales de la serie, los valores predichos, los errores y los intervalos de confianza para las predicciones (Figura 11-58).

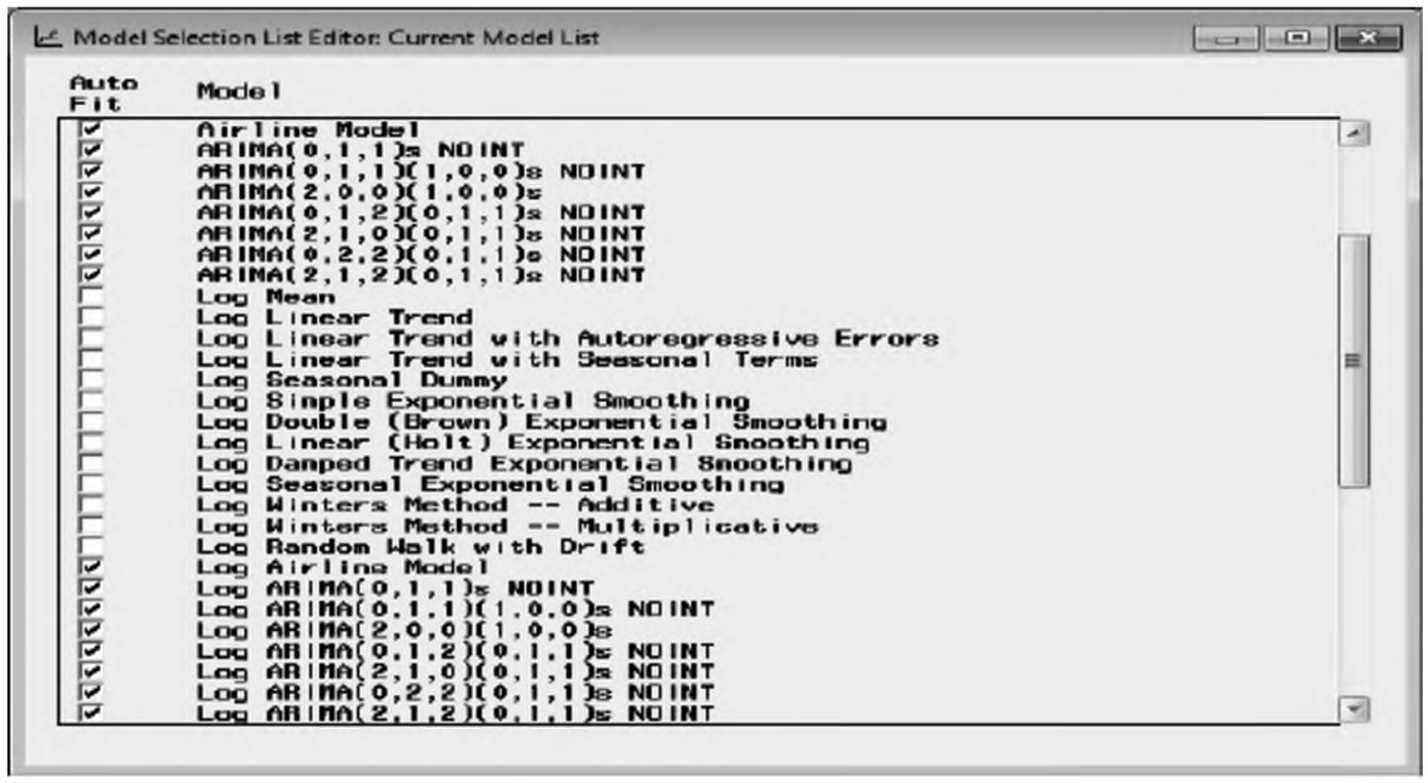


Figura 11-52

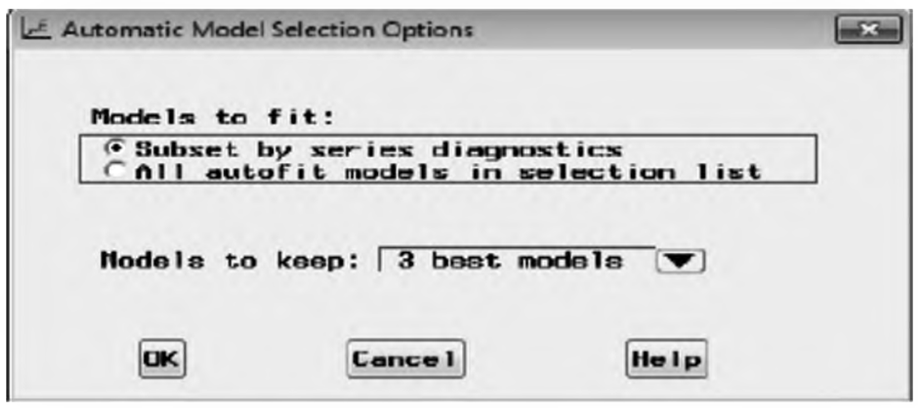


Figura 11-53

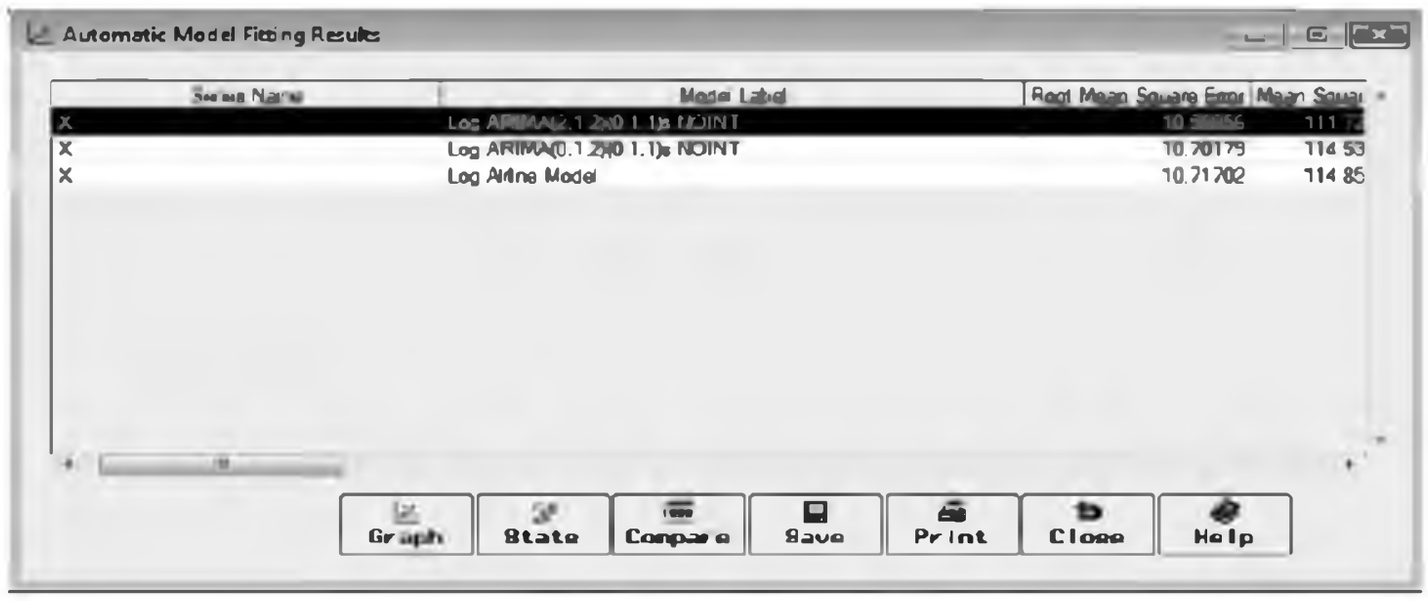


Figura 11-54

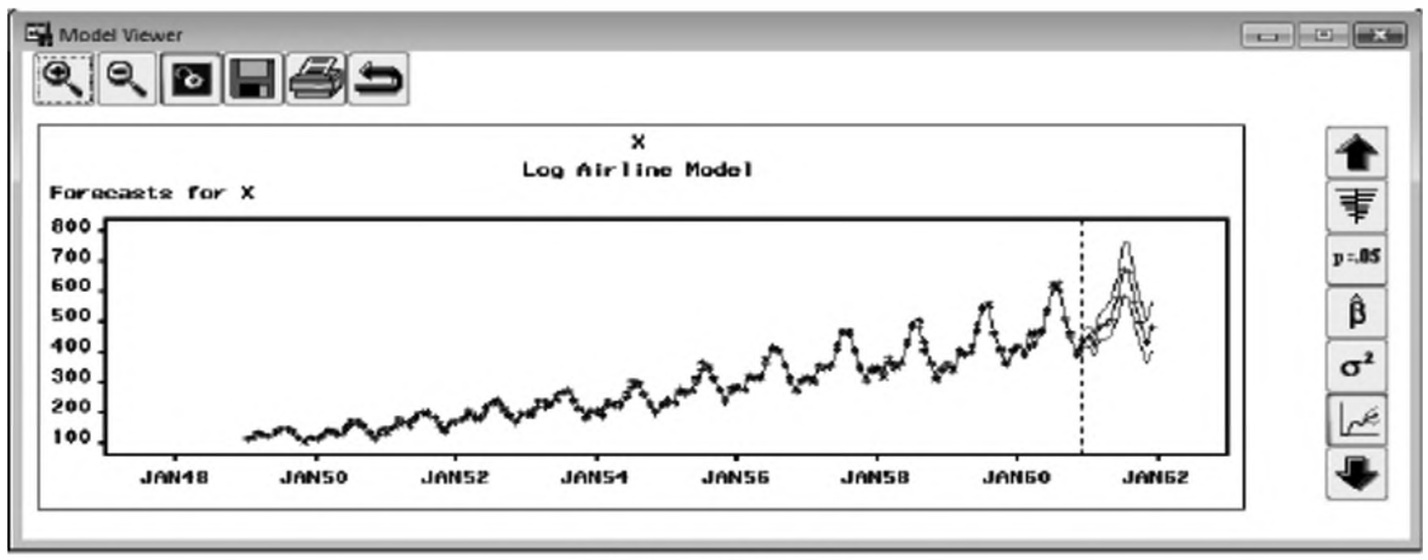


Figura 11-55

©Alfaomega-RC Libros

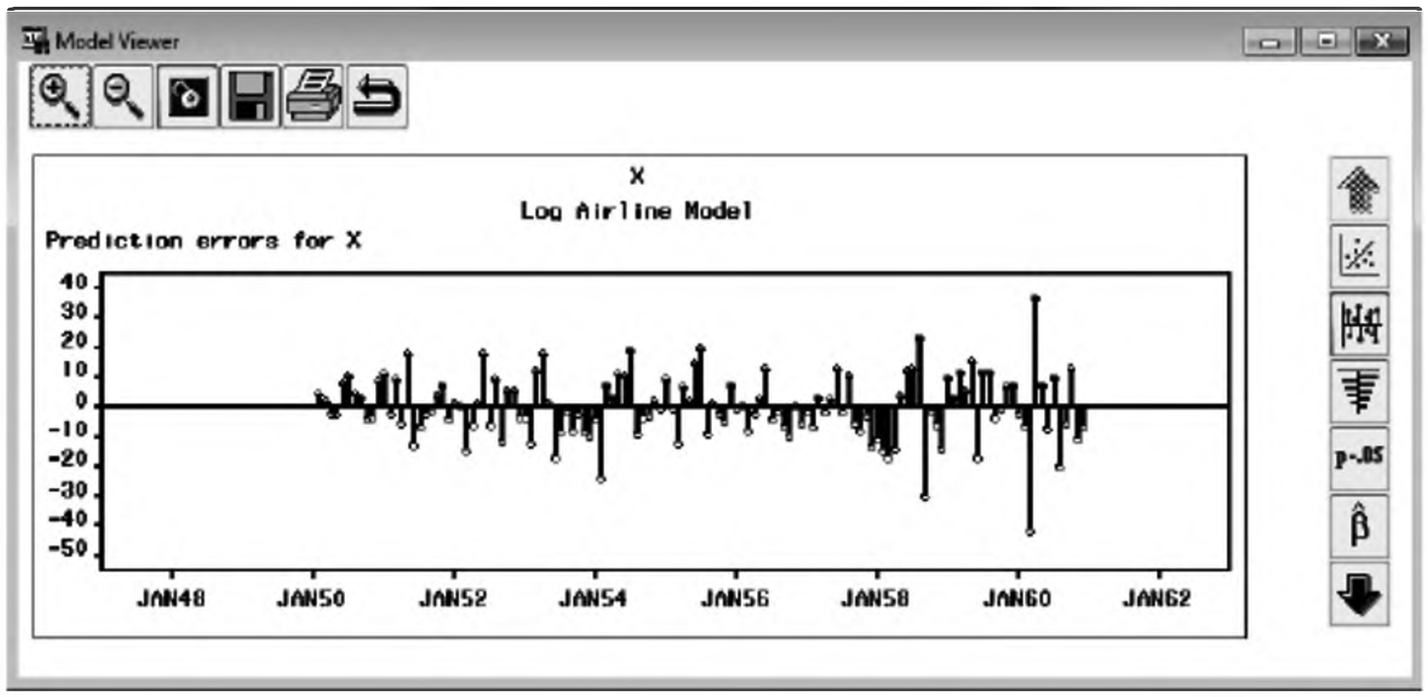
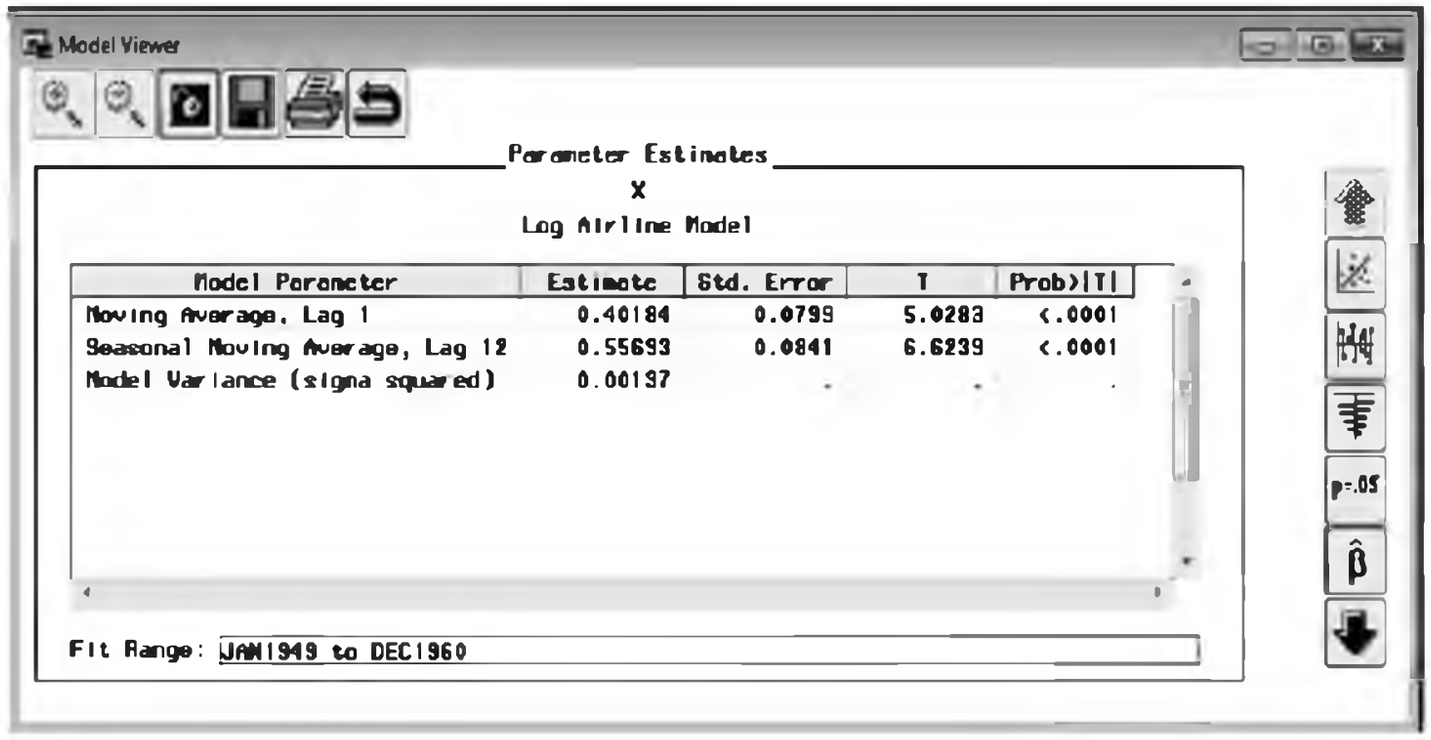
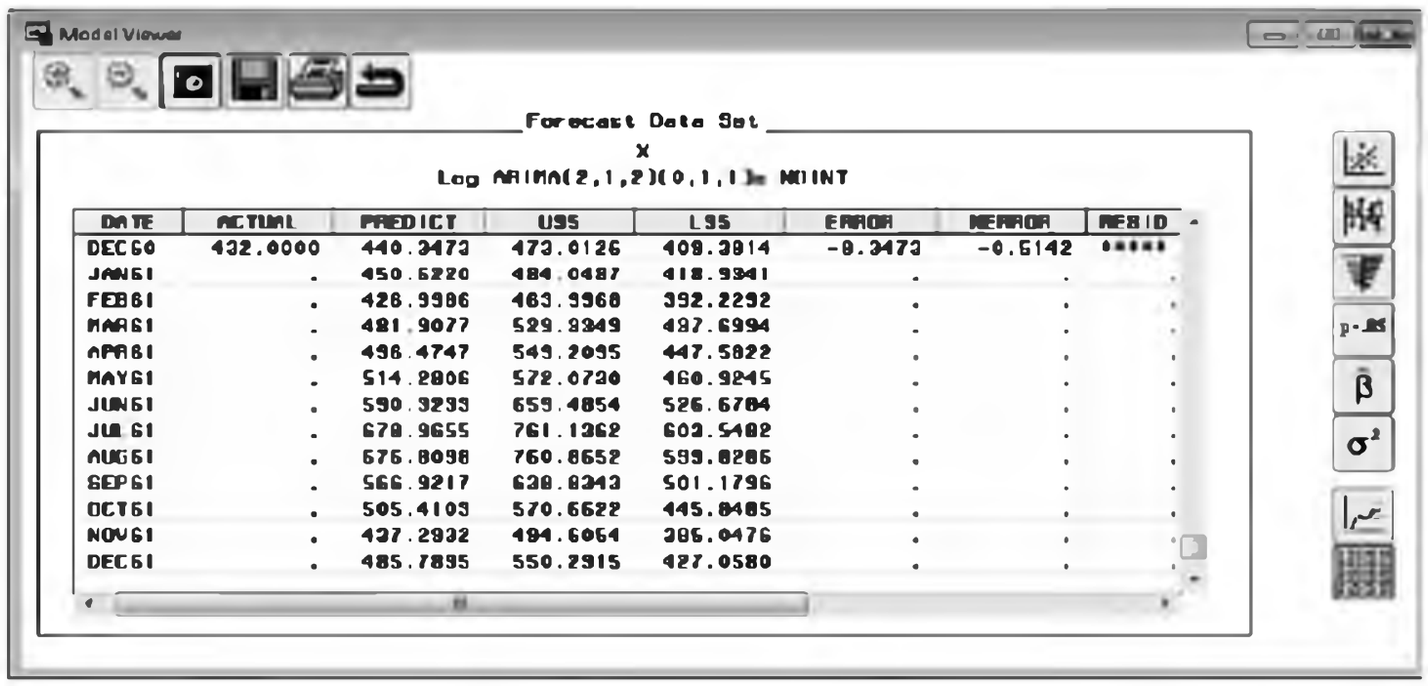


Figura 11-56



Model Parameter	Estimate	Std. Error	T	Prob>|T|
Moving Average, Lag 1	0.40184	0.0799	5.0283	<.0001
Seasonal Moving Average, Lag 12	0.55693	0.0841	6.6239	<.0001
Model Variance (sigma squared)	0.00197	.	.	.

Figura 11-57



DATE	ACTUAL	PREDICT	U95	L95	ERROR	NERROR	RESID
DEC60	432.0000	440.3473	473.0126	408.3814	-8.3473	-0.5142	*****
JAN61	.	450.6220	484.0487	418.9941	.	.	.
FEB61	.	426.9986	463.9968	392.2292	.	.	.
MAR61	.	481.9077	529.9349	437.6994	.	.	.
APR61	.	498.4747	549.2095	447.5822	.	.	.
MAY61	.	514.2806	572.0730	460.9245	.	.	.
JUN61	.	590.3233	659.4854	526.6784	.	.	.
JUL61	.	678.9655	761.1362	602.5482	.	.	.
AUG61	.	676.8098	760.8652	593.8286	.	.	.
SEP61	.	566.9217	638.8343	501.1796	.	.	.
OCT61	.	505.4103	570.6622	445.8485	.	.	.
NOV61	.	437.2932	494.6064	386.0476	.	.	.
DEC61	.	485.7895	550.2915	427.0580	.	.	.

Figura 11-58

CAPÍTULO 12

MODELOS AUTOMÁTICOS Y EVALUACIÓN DE MODELOS CON HERRAMIENTAS DE IBM Y SAS

MODELIZACIÓN AUTOMÁTICA EN IBM SPSS MODELER

Dentro de la minería de datos desempeñan un papel muy importante los algoritmos de modelización automática. Estas herramientas permiten elegir el mejor modelo que se ajusta a los datos para realizar tareas predictivas óptimas con suficiente criterio. Estas herramientas pueden encuadrarse incluso en la fase de evaluación de la minería de datos, ya que la selección del modelo óptimo se produce después de un proceso de evaluación que decide el mejor algoritmo.

IBM SPSS incorpora los algoritmos de modelización automática en la categoría *Automático* de la ficha *Modelado* (Figura 12-1). En concreto se dispone de los algoritmos: *Clasificador automático* (para elegir el mejor modelo predictivo con variable dependiente categórica), *Autonumérico* (para elegir el mejor modelo predictivo con variable dependiente cuantitativa), *Autoconglomerado* (para elegir la mejor técnica de segmentación cluster) y *Serie temporal* (para elegir el mejor modelo de predicción con series temporales).

Figura 12-1

EL NODO CLASIFICADOR AUTOMÁTICO DE IBM SPSS MODELER

El nodo *Clasificador automático* crea y compara varios modelos diferentes para obtener resultados binarios (sí o no, pérdida o no de clientes, etc.), lo que le permite seleccionar el mejor enfoque para un análisis determinado. Son compatibles varios algoritmos de modelado, por lo que es posible seleccionar los métodos que desee utilizar, las opciones específicas para cada uno y los criterios para comparar los resultados. El nodo genera un conjunto de modelos basado en las opciones especificadas y clasifica los mejores candidatos en función de los criterios que especifique.

Puede usar la configuración por defecto o seleccionar las opciones para cada tipo de modelo automático. Las opciones específicas son parecidas a las disponibles en los nodos de modelado independientes, con la diferencia de que en lugar de elegir un ajuste u otro, puede seleccionar todos los que desee aplicar en la mayoría de los casos. Por ejemplo, si compara modelos Red neuronal, puede seleccionar varios métodos de entrenamiento diferentes y probar cada método con semilla aleatoria y sin ella. Se utilizarán todas las combinaciones posibles de las opciones seleccionadas, facilitando la generación de muchos modelos diferentes de una única pasada. No obstante, tenga cuidado, ya que la selección de varios ajustes puede hacer que el número de modelos se multiplique muy rápidamente.

En general, para seleccionar las opciones para cada tipo de modelo:

- En el nodo de modelado automatizado, seleccione la pestaña *Experto*.
- Haga clic en la columna *Parámetros de modelo* para el tipo de modelo.
- En el menú desplegable, seleccione *Especificar*.
- En el cuadro de diálogo *Configuración de algoritmo*, seleccione las opciones de la columna *Opciones*.

 ©Alfaomega-RC Libros

Las *reglas de parada* especificadas para los nodos de modelado automático están relacionadas con la ejecución global del nodo, no con la parada de modelos determinados generados por el nodo. Se tendrá en cuenta lo siguiente:

- *Limitar tiempo de ejecución global a*. (Solo modelos de Red neuronal, K-medias, Kohonen, Bietápico, SVM, KNN, Red bayesiana y C&R Tree). Detiene la ejecución tras un número específico de horas. Se incluirán en el nugget de modelo todos los modelos generados hasta ese momento, pero no se producirán más modelos.

- *Deténgalo en cuanto se produzcan modelos válidos*. Detiene la ejecución cuando un modelo cumple todos los criterios especificados en la pestaña *Descartar* (para el nodo Clasificador automático o Autoconglomerado) o la pestaña *Modelo* (para el nodo Autonumérico).

El **nodo Clasificador automático** calcula y compara los modelos de objetivos (variables dependientes) nominales (conjuntos) o binarios (yes/no) utilizando métodos diferentes y permitiéndole probar una gran variedad de métodos en una única tirada. Puede seleccionar los algoritmos que se utilizarán y experimentar con múltiples combinaciones de opciones. Por ejemplo, en lugar de elegir entre los métodos rápido, dinámico o de poda de una red neuronal, puede probarlos todos. El nodo explora cada combinación posible de opciones, evalúa el modelo de cada candidato en función de la medida especificada y guarda los mejores para su uso en la puntuación o en futuros análisis.

Se requiere un campo de objetivo con un nivel de medición de *Nominal* o *Marca* (con el papel establecido a *Objetivo*), y al menos un campo de entrada (con el papel establecido a *Entrada*). En un campo de marca, el valor *Verdadero* definido para el campo objetivo se supone que representa un acierto al calcular beneficios, elevación y estadísticos relacionados. Los campos de entrada pueden tener un nivel de medición de *Continuo* o *Categórico*, con la limitación de que algunas entradas pueden no ser apropiadas para algunos tipos de modelos. Por ejemplo, los campos ordinales que se utilizan como entradas en los modelos C&RT, CHAID y QUEST deben tener almacenamiento numérico (no en cadenas). Asimismo, estos modelos los omitirán si se especifica lo contrario. De igual modo, los campos de entrada continuos pueden establecerse en intervalos en algunos casos. Los requisitos son los mismos que cuando se utilizan los nodos de modelado individuales. Por ejemplo, un modelo Red bayesiana funciona igual independientemente de si se ha generado desde el nodo Red bayesiana o el nodo Clasificador automático.

Pueden utilizarse *Campos de frecuencia y ponderación*. La frecuencia y la ponderación se usan para proporcionar importancia adicional a ciertos registros sobre otros porque, por ejemplo, el usuario sabe que el conjunto de datos creado no representa totalmente una sección de la población principal (*Ponderación*) o porque un registro representa un número de casos idénticos (*Frecuencia*). Si se especifica, los modelos C&RT, CHAID, QUEST, Lista de decisiones y Red bayesiana pueden utilizar un campo de frecuencia. Los modelos C&RT, CHAID y C5.0 pueden utilizar un campo de ponderación. Otros tipos de modelo omitirán estos campos y crearán los modelos de todas formas. Los campos de frecuencia y ponderación solo se utilizan para la creación de modelos y no se tienen en cuenta al evaluar o puntuar modelos.

Los **tipos de modelo admitidos** incluyen Red neuronal, C&RT, QUEST, CHAID, C5.0, Regresión logística, Lista de decisiones, Red bayesiana, Discriminante, Vecino más cercano y SVM.

Como ejemplo consideramos un modelo que trata de predecir la probabilidad de que un cliente de un banco devuelva un crédito (variable CREDIT_V) conocidos los valores de ciertas variables que determinan sus características de edad (EDAD), sueldo (PAGO_MES), categoría profesional (CAT_PROF) y disponibilidad de tarjetas de crédito (AMEX). La variable CREDIT_V será la variable objetivo (dependiente) y el resto de las variables serán las variables de entrada del modelo (independientes).

Una vez importado el fichero como un origen de datos ASCII, utilizar el nodo *Tipo* para asignar a la variable *CREDIT_V* el papel de variable *objetivo* y al resto de las variables el papel de *entrada* salvo el número de cliente (Figura 12-2). A continuación se une al diagrama el nodo *Clasificador automático*, se hace doble clic sobre él con el botón secundario del ratón y se elige *Edición* en el menú emergente resultante para definir los parámetros en la Figura 12-3. En la solapa *Campos* elegimos *Utilizar configuración del nodo Tipo*. En la solapa *Modelo* (Figura 12-4) definimos las características generales de los modelos a ajustar. En la solapa *Experto* (Figura 12-5) se seleccionan las opciones para cada tipo de modelo, en la solapa *Descartar* (Figura 12-6) se definen condiciones de exclusión de modelos en el análisis. Al hacer clic en *Aplicar* y *Ejecutar* se obtienen las pantallas temporales de las Figuras 12-7 y 12-8 que informan del proceso de ejecución. Finalmente se obtiene el modelo sobre la ruta (Figura 12-9).

©Alfaomega-RC Libros

Figura 12-2

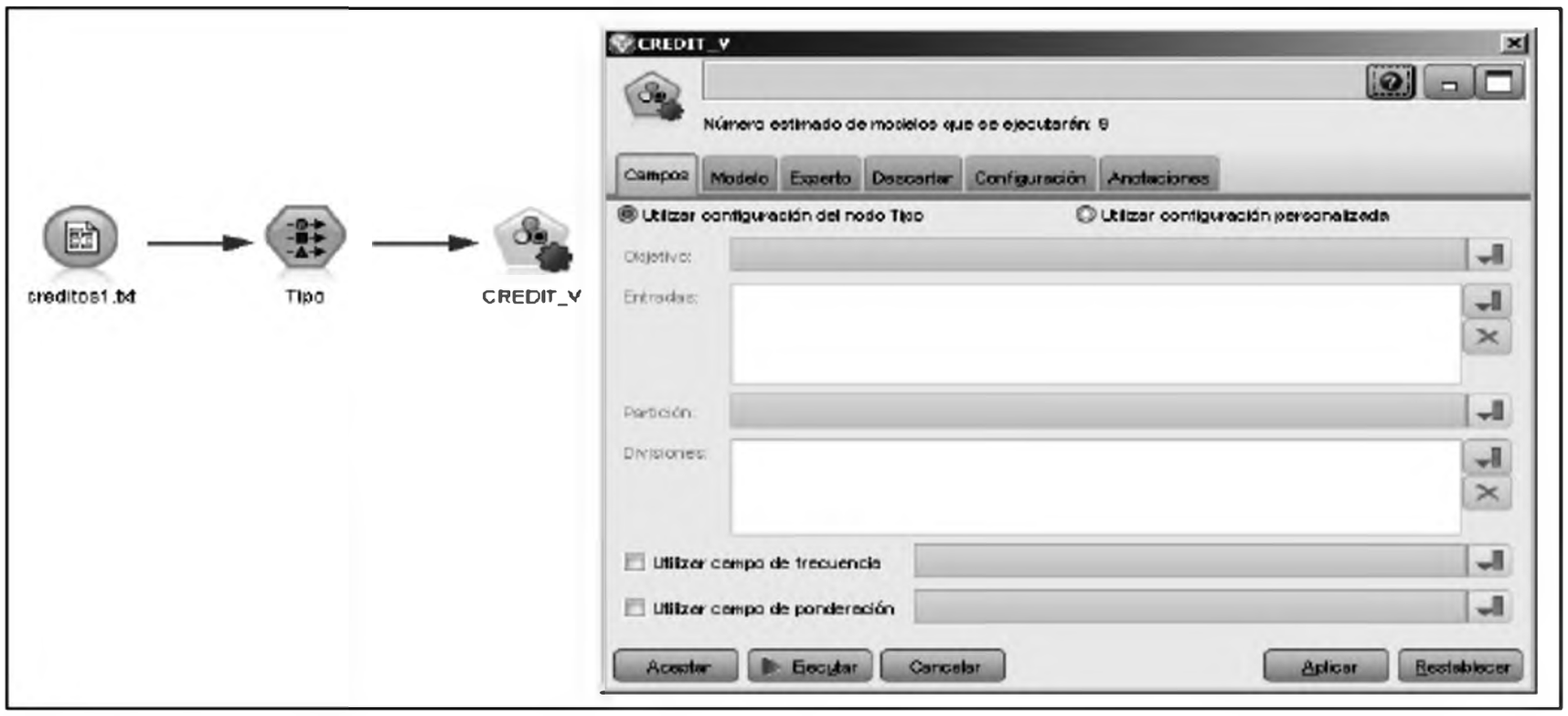


Figura 12-3

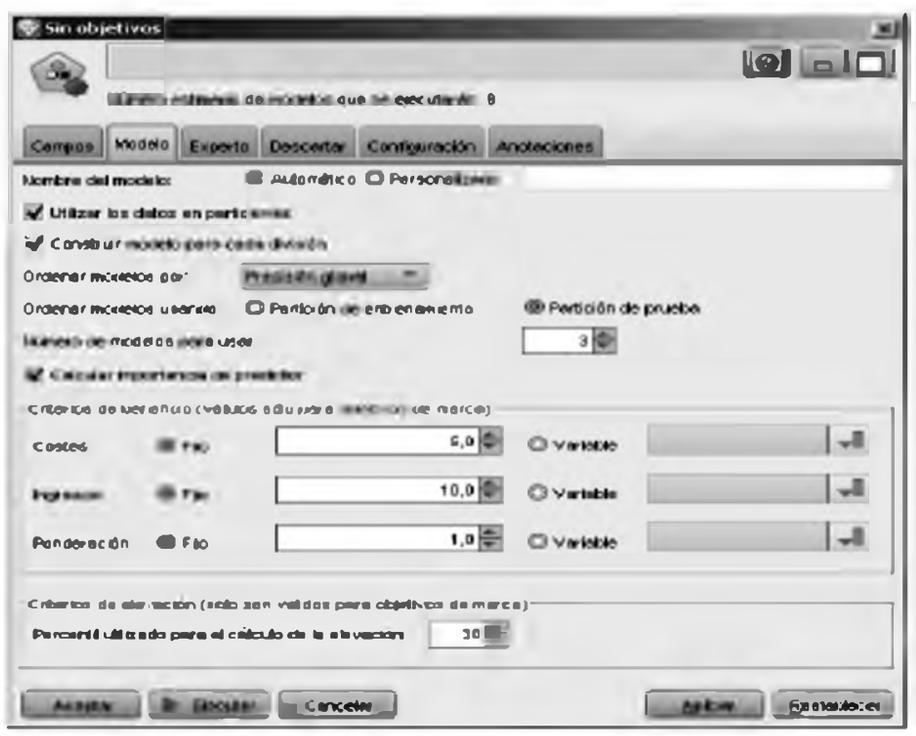


Figura 12-4

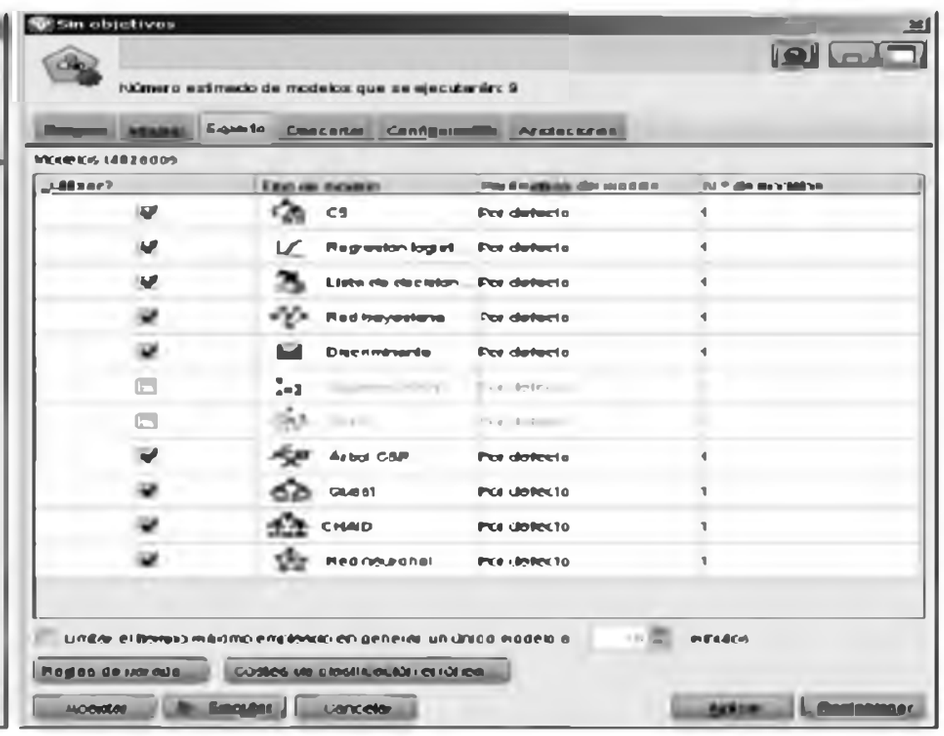


Figura 12-5

Figura 12-6

Figura 12-7

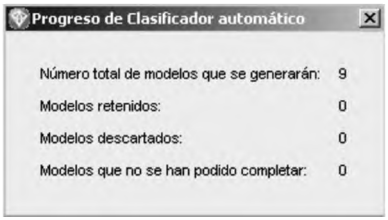


Figura 12-8

La pestaña *Modelo* del nodo *Clasificador automático* (Figura 12-4) le permite especificar el número de modelos que se van a crear, junto con los criterios empleados para compararlos. En esta pestaña tenemos los siguientes campos:

Nombre del modelo. Puede generar el nombre del modelo de forma automática basándose en el campo objetivo o de ID (o en el nombre del tipo de modelo si se especifica ningún campo objetivo), o bien especificar un nombre personalizado.

Utilizar los datos en particiones. Si se ha definido un campo de partición, esta opción garantiza que solo se utilizarán los datos de la partición de entrenamiento para la generación del modelo. Consulte el tema Nodo Partición si desea obtener más información.

Crear modelos divididos. Genera un modelo diferente para cada valor posible de campos de entrada que se especifican como modelos divididos. Consulte el tema Generación de modelos divididos si desea obtener más información.

Ordenar modelos por. Especifique los criterios utilizados para comparar y clasificar los modelos. Las opciones incluyen la precisión global, el área debajo de la curva ROC,

 ©Alfaomega-RC Libros

beneficio, elevación y número de campos. Tenga en cuenta que todas estas medidas estarán disponibles en el informe de resumen independientemente de lo que se seleccione aquí.

Nota: En el caso de un objetivo nominal (conjunto), la ordenación está restringida a Precisión global o Número de campos.

Al calcular beneficios, elevación y estadísticos relacionados, se supone que el valor *True* definido para el campo objetivo representa un acierto.

Ordenar modelos usando. Si se está usando una partición, puede especificar si los rangos se basan en el conjunto de datos de entrenamiento o en el conjunto de prueba. En conjuntos de datos de gran tamaño, si usa una partición para el filtrado preliminar de modelos, puede mejorar rendimiento en gran medida.

Número de modelos que se utilizarán. Especifica el número máximo de modelos que aparecerán en el nugget de modelo generado por el nodo. Los primeros modelos de la lista se enumeran en función del criterio de ordenación especificado. Tenga en cuenta que si aumenta este límite puede ralentizarse el rendimiento. El valor máximo permitido es 100.

Calcular importancia del predictor. En el caso de modelos que produzcan una medida adecuada de importancia, puede mostrar un gráfico que indique la importancia relativa de cada predictor al estimar el modelo. Normalmente, desea centrar sus esfuerzos de modelado en los predictores que importan más y considera eliminar o ignorar los que importan menos. Tenga en cuenta que la importancia de predictor puede aumentar el tiempo necesario para calcular algunos modelos; además no se recomienda si solo desea una amplia comparación entre varios modelos diferentes. Es de mayor utilidad una vez ha limitado su análisis a unos cuantos modelos que desee explorar más a fondo. Consulte el tema Importancia del predictor si desea obtener más información.

Criterios de beneficio. Solo para objetivos de marca. El beneficio es igual a los ingresos de cada registro menos el coste del registro. Los beneficios de un cuantil son la suma de los beneficios de todos los registros del cuantil. Se asume que los beneficios se aplican solo a los aciertos, pero los costes se aplican a todos los registros.

- *Costes*. Permite especificar el coste asociado con cada registro. Puede seleccionar costes fijos o variables. En el caso de los costes fijos, especifique el valor del coste. En el caso de los costes variables, pulse en el selector de campos para elegir un campo de costes.

• *Ingresos*. Permite especificar los ingresos asociados con cada registro que representa un acierto. Puede seleccionar costes fijos o variables. En el caso de los ingresos fijos, especifique el valor del ingreso. En el caso de los ingresos variables, pulse en el selector de campos para elegir un campo de ingresos.

• *Ponderación*. Si los registros de los datos representan más de una unidad, puede utilizar ponderaciones de frecuencias para ajustar los resultados. Especifique la ponderación asociada con cada registro mediante valores fijos o variables. En el caso de las ponderaciones fijas, especifique el valor de ponderación (el número de unidades por registro). En el caso de ponderaciones variables, pulse en el selector de campos para elegir un campo de ponderaciones.

Criterios de elevación. Solo para objetivos de marca. Especifica el percentil que hay que utilizar para los cálculos de la elevación. Tenga en cuenta que también puede cambiar este valor al comparar los resultados. Consulte el tema Nugget de modelo automático si desea obtener más información.

La pantalla *Experto* (Figura 12-5) define las opciones de experto para el nodo Clasificador automático. Esta pestaña le permite aplicar una partición (si está disponible), seleccionar los algoritmos que se van a usar y especificar las reglas de parada. Las opciones son las siguientes:

Modelos utilizados. Use las casillas de verificación de la columna izquierda para seleccionar los tipos de modelo (algoritmos) que se van a incluir en la comparación. Cuantos más tipos seleccione, más modelos se crearán y más tardará el procesamiento.

Tipo de modelo. Enumera los algoritmos disponibles (consulte a continuación).

Parámetros del modelo. Puede usar la configuración por defecto o seleccionar *Especificar* para elegir opciones para cada tipo de modelo. Las opciones específicas son parecidas a las disponibles en los nodos de modelado independientes, con la diferencia de que se pueden seleccionar varias opciones o combinaciones. Por ejemplo, si compara los modelos del nodo Red neuronal, puede seleccionar los seis modelos para entrenarlos de una vez en lugar de seleccionar uno de ellos.

Número de modelos. Enumera el número de modelos generados para cada algoritmo basados en la configuración actual. Al combinar opciones, puede aumentar rápidamente el número de modelos, por lo que se recomienda prestar especial atención a este número, especialmente si usa conjuntos de datos grandes.

 ©Alfaomega-RC Libros

Limitar el tiempo máximo empleado en generar un único modelo. (Solo modelos de K-medias, Kohonen, bietápicos, SVM, KNN, de red bayesiana y de lista de decisiones) Establece un límite de tiempo máximo para cualquier modelo. Por ejemplo, si un modelo determinado necesita un período de tiempo más largo del esperado para entrenarse debido a una interacción compleja, es probable que no quiera detener la ejecución de todo el modelado.

Nota: si el objetivo es un campo nominal (conjunto), la opción *Lista de decisiones* no está disponible.

Los algoritmos admitidos son los siguientes:

Red neuronal. El nodo Red neuronal utiliza un modelo simplificado que emula el modo en que el cerebro humano procesa la información. Funciona simultaneando un número elevado de unidades simples de procesamiento interconectadas que parecen versiones abstractas de neuronas. Las redes neuronales son estimadores potentes de funciones generales y requieren un conocimiento matemático o estadístico mínimo para entrenarlas o aplicarlas.

Nodo C5.0. El nodo C5.0 genera un árbol de decisión o un conjunto de reglas. El modelo divide la muestra basándose en el campo que ofrece la máxima ganancia de información en cada nivel. El campo objetivo debe ser categórico. Se permiten varias divisiones en más de dos subgrupos.

Nodo C&RT. El nodo de árbol de clasificación y regresión (C&R) genera un árbol de decisión que permite pronosticar o clasificar observaciones futuras. El método utiliza la partición reiterada para dividir los registros de entrenamiento en segmentos minimizando las impurezas en cada paso, donde un nodo se considera "puro" si el 100% de los casos del nodo corresponden a una categoría específica del campo objetivo. Los campos de entrada y objetivo pueden ser continuos (rango numérico) o categóricos (nominal, ordinal o marca). Todas las divisiones son binarias (solo se crean dos subgrupos).

Nodo QUEST. Proporciona un método de clasificación binario para generar árboles de decisión; está diseñado para reducir el tiempo de procesamiento necesario para realizar los análisis de C&RT y reducir la tendencia de los métodos de clasificación de árboles para favorecer a las entradas que permitan realizar más divisiones. Los campos de entrada pueden ser continuos (rango numérico); sin embargo, el campo objetivo debe ser categórico. Todas las divisiones son binarias.

Nodo CHAID. El nodo CHAID genera árboles de decisión utilizando estadísticos de chi-cuadrado para identificar las divisiones óptimas. A diferencia de los nodos C&RT y QUEST, CHAID puede generar árboles no binarios, lo que significa que algunas divisiones generarán más de dos ramas. Los campos de entrada y objetivo pueden ser continuos (rango numérico) o categóricos. CHAID exhaustivo es una modificación de CHAID que examina con mayor precisión todas las divisiones posibles, aunque necesita más tiempo para realizar los cálculos.

Nodo Logística. La regresión logística es una técnica de estadístico para clasificar los registros en función de los valores de los campos de entrada. Es análoga a la regresión lineal pero toma un campo objetivo categórico en lugar de uno numérico.

Nodo Lista de decisiones. El nodo Lista de decisiones identifica subgrupos, o segmentos, que muestran una mayor o menor posibilidad de proporcionar un resultado binario relacionado con la población global. Por ejemplo, puede buscar clientes que tengan menos posibilidades de perder clientes o más posibilidades de responder favorablemente a una campaña. Puede incorporar su conocimiento empresarial al modelo añadiendo sus propios segmentos personalizados y previsualizando modelos alternativos uno junto a otro para comparar los resultados. Los modelos de listas de decisiones constan de una lista de reglas en las que cada regla tiene una condición y un resultado. Las reglas se aplican en orden, y la primera regla que coincide determina el resultado.

Nodo Red bayesiana. Permite crear un modelo de probabilidad combinando pruebas observadas y registradas con conocimiento del mundo real para establecer la probabilidad de instancias. El nodo se centra en las redes Naïve Bayes aumentado a árbol (TAN) y de cadena de Markov que se utilizan principalmente para la clasificación.

Nodo Análisis discriminante. El análisis discriminante realiza más supuestos rigurosos que regresiones logísticas, pero puede ser una alternativa o un suplemento valioso al análisis de regresión logística si se cumplen dichos supuestos. Consulte el tema Nodo Discriminante si desea obtener más información.

Nodo KNN. El nodo k de modelado de vecino (KNN) asocia el nuevo caso con la categoría o valor de los objetos k junto a él en el espacio de predictores, donde k es un entero. Los casos parecidos están próximos y los que no lo son están alejados entre sí.

 ©Alfaomega-RC Libros

Nodo SVM. El nodo Máquina de vectores de soporte (SVM) le permite clasificar datos en uno o dos grupos sin que haya un ajuste por exceso. SVM funciona bien con conjuntos de datos grandes, como aquellos con un gran número de campos de entrada.

La pestaña *Descartar* (Figura 12-6) del nodo Clasificador automático le permite descartar automáticamente los modelos que no cumplen determinados criterios. Estos modelos no aparecerán enumerados en el informe de resumen. Puede especificar un umbral mínimo para la precisión global y un umbral máximo para el número de variables usadas en el modelo. Además, en el caso de objetivos de marca, puede especificar un umbral mínimo para la elevación, los beneficios y un área debajo de la curva; la elevación y los beneficios se determinan según lo especificado en la pestaña *Modelo*. Si lo desea, puede configurar el nodo para que se detenga la ejecución la primera vez que se genere un modelo que cumpla todos los criterios especificados.

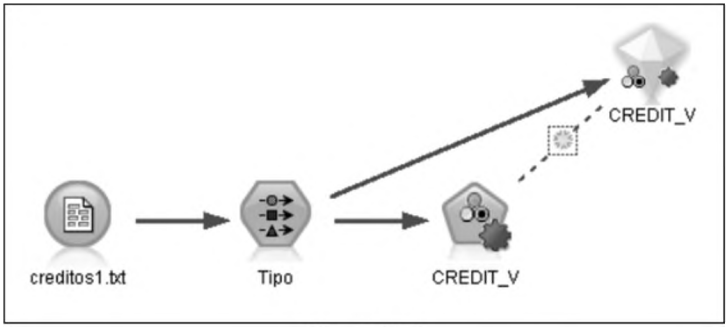


Figura 12-9

Si ahora hacemos clic con el botón derecho del ratón en el nodo *Clasificador automático* sobre la ruta, se obtiene la pantalla de resultados de la Figura 12-10. En la solapa *Modelo* de esta figura se observa que el mejor modelo predictivo estimado es el árbol C&R, seguido del árbol CHAID y del modelo de Red neuronal. En las diferentes columnas de la Figura 12-10 se observan los parámetros de los modelos seleccionados. En la solapa *Gráfico* (Figura 12-11) se observan los gráficos de importancia del predictor y de distribución de las categorías de la variable predicha para los modelos seleccionados.

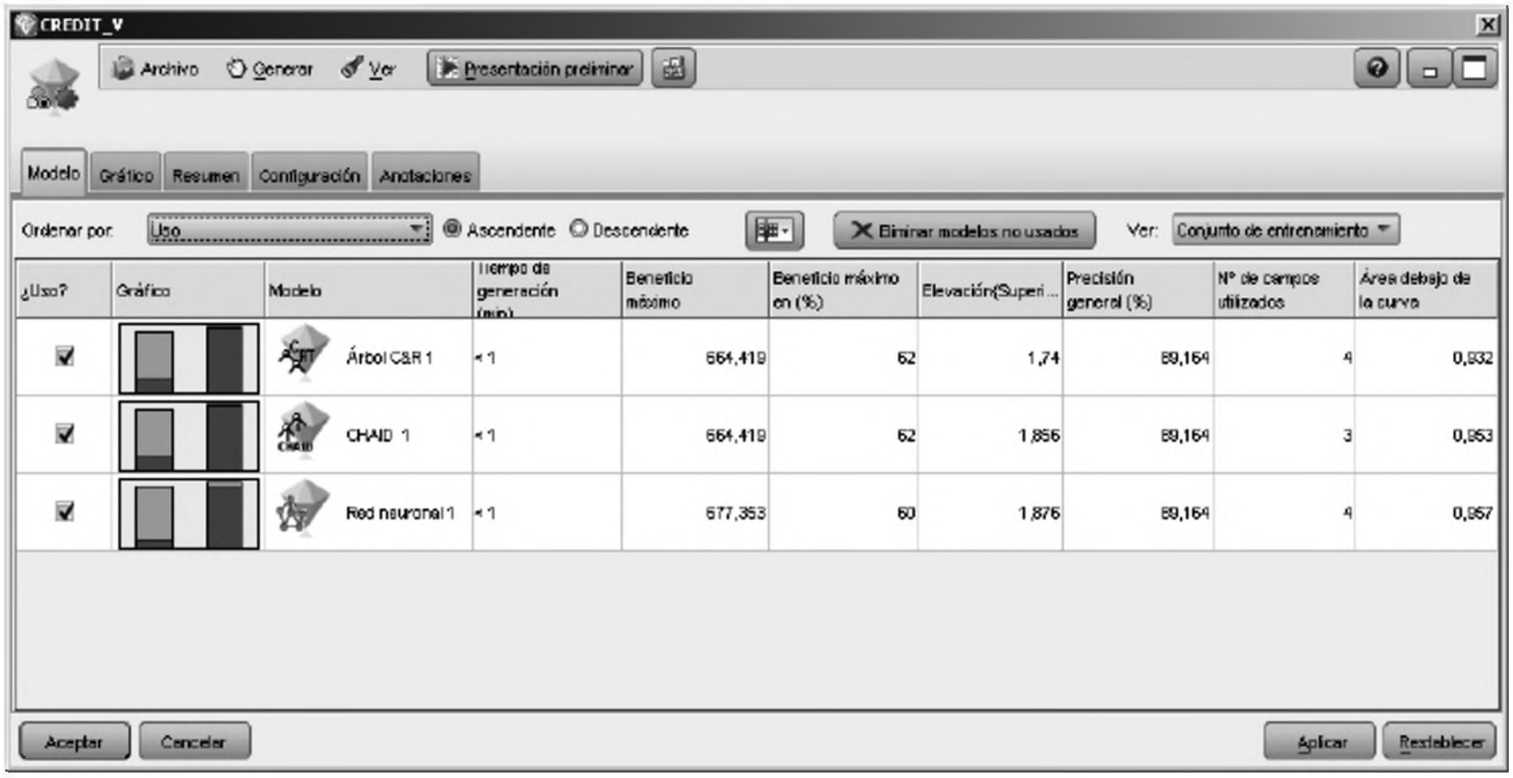


Figura 12-10

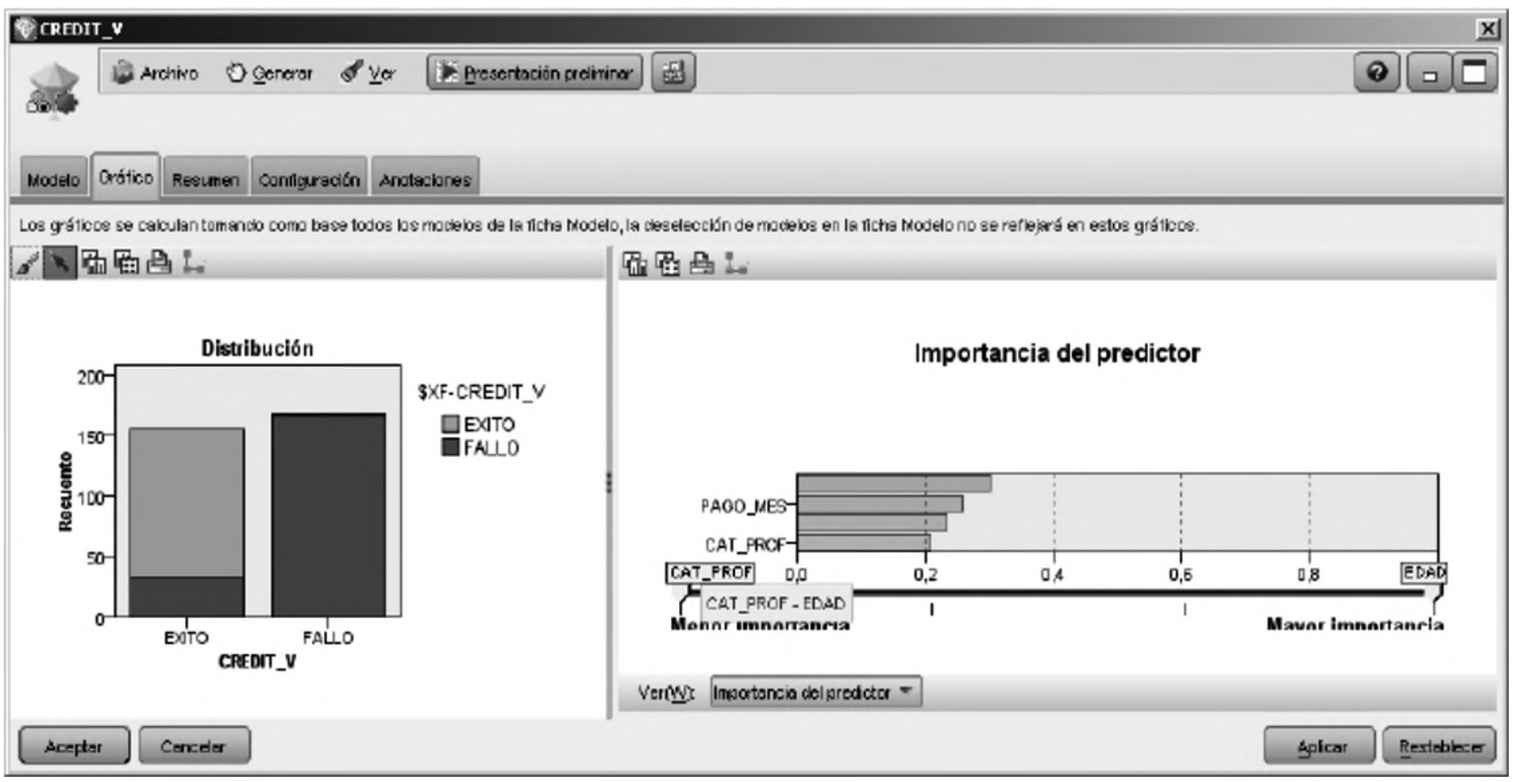
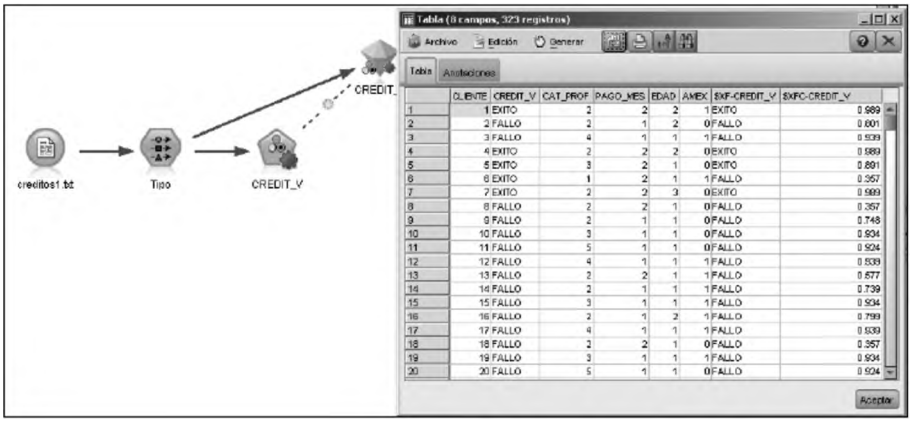


Figura 12-11

Si se añade un nodo *Tabla* al nodo *Clasificador automático* (Figura 12-12) se observan dos nuevos campos que contienen la pertenencia de cada registro (cliente) a la categoría correspondiente de la variable dependiente y la probabilidad de asignación del cliente a esa categoría. Es decir, conocemos si el cliente devolverá o no el crédito y con qué probabilidad.

©Alfaomega-RC Libros



	CLIENTE	CREDIT_V	CAT_PROF	PAGO_MES	EDAD	AMEX	$XF-CREDIT_V	$XFC-CREDIT_V
1	1	EXITO	2	2	2	1	EXITO	0.989
2	2	FALLO	2	1	2	0	FALLO	0.601
3	3	FALLO	4	1	1	1	FALLO	0.539
4	4	EXITO	2	2	2	0	EXITO	0.989
5	5	EXITO	3	2	1	0	EXITO	0.891
6	6	EXITO	1	2	1	1	FALLO	0.357
7	7	EXITO	2	2	3	0	EXITO	0.989
8	8	FALLO	2	2	1	0	FALLO	0.357
9	9	FALLO	2	1	1	0	FALLO	0.748
10	10	FALLO	3	1	1	0	FALLO	0.934
11	11	FALLO	5	1	1	0	FALLO	0.924
12	12	FALLO	4	1	1	1	FALLO	0.539
13	13	FALLO	2	2	1	1	FALLO	0.577
14	14	FALLO	2	1	1	1	FALLO	0.739
15	15	FALLO	3	1	1	1	FALLO	0.934
16	16	FALLO	2	1	2	1	FALLO	0.799
17	17	FALLO	4	1	1	1	FALLO	0.539
18	18	FALLO	2	2	1	0	FALLO	0.357
19	19	FALLO	3	1	1	1	FALLO	0.934
20	20	FALLO	5	1	1	0	FALLO	0.924

Figura 12-12

EL ALGORITMO AUTONUMÉRICO DE IBM SPSS MODELER

El nodo Autonumérico calcula y compara modelos para resultados de rango numérico continuo utilizando cierto número de métodos diferentes. El nodo funciona de la misma manera que el nodo Clasificador automático, lo que le permite seleccionar los algoritmos que desee utilizar y experimentar con varias combinaciones de opciones en una única pasada de modelado. Los algoritmos admitidos incluyen redes neuronales, C&RT, CHAID, regresión lineal, regresión lineal generalizada y máquinas de vectores de soporte (SVM). Los modelos se pueden comparar basándose en la correlación, el error relativo o el número de variables utilizado.

El nodo Autonumérico calcula y compara los modelos de resultados de rango numérico continuo utilizando métodos diferentes y permitiéndole probar una gran variedad de métodos en una única tirada. Puede seleccionar los algoritmos que se utilizarán y experimentar con múltiples combinaciones de opciones. Por ejemplo, puede pronosticar valores de viviendas utilizando los modelos Red neuronal, Regresión lineal, C&RT y CHAID para ver cuál tiene el mejor rendimiento; asimismo, puede probar diferentes combinaciones de métodos de regresión Por pasos, Adelante y Hacia atrás. El nodo explora cada combinación posible de opciones, evalúa el modelo de cada candidato en función de la medida especificada y guarda los mejores para su uso en la puntuación o en futuros análisis.

Como ejemplo consideramos un modelo que trata de predecir el consumo de gasolina de los automóviles en función de la potencia, peso, motor, aceleración,

origen y cilindrada de los vehículos. La variable consumo será la variable objetivo (dependiente) y el resto de las variables serán las variables de entrada del modelo (independientes).

Una vez importado el fichero *Coches.sav* como un origen de datos IBM Statistics, utilizar el nodo *Tipo* para asignar a la variable *consumo* el papel de variable *objetivo* y al resto de las variables el papel de *entrada* salvo el año y una variable derivada que hay en el archivo (Figura 12-13). A continuación se une al diagrama el nodo *Autonumérico*, se hace doble clic sobre él con el botón secundario del ratón y se elige *Edición* en el menú emergente resultante para definir los parámetros en la Figura 12-14. En la solapa *Campos* elegimos *Utilizar configuración del nodo Tipo*. En la solapa *Modelo* (Figura 12-15) definimos las características generales de los modelos a ajustar. En la solapa *Experto* (Figura 12-16) se seleccionan las opciones para cada tipo de modelo. Al hacer clic en *Aplicar* y *Ejecutar* se obtiene el modelo sobre la ruta (Figura 12-17).

Figura 12-13

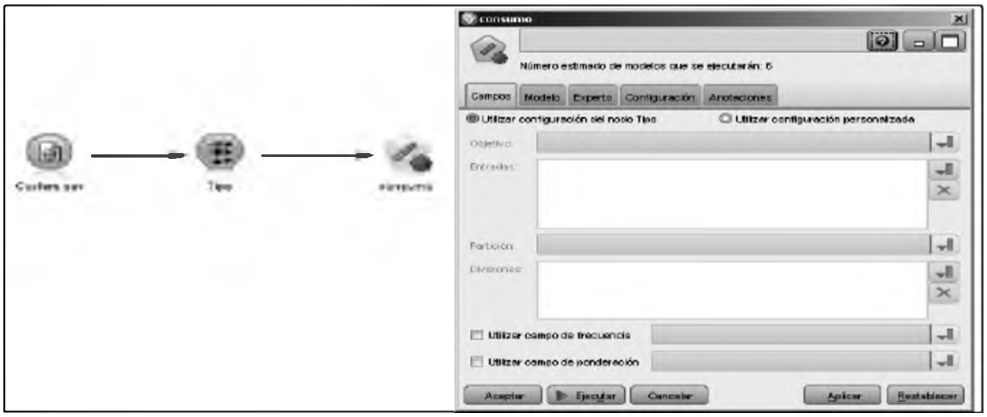


Figura 12-14

©Alfaomega-RC Libros

Figura 12-15

Figura 12-16

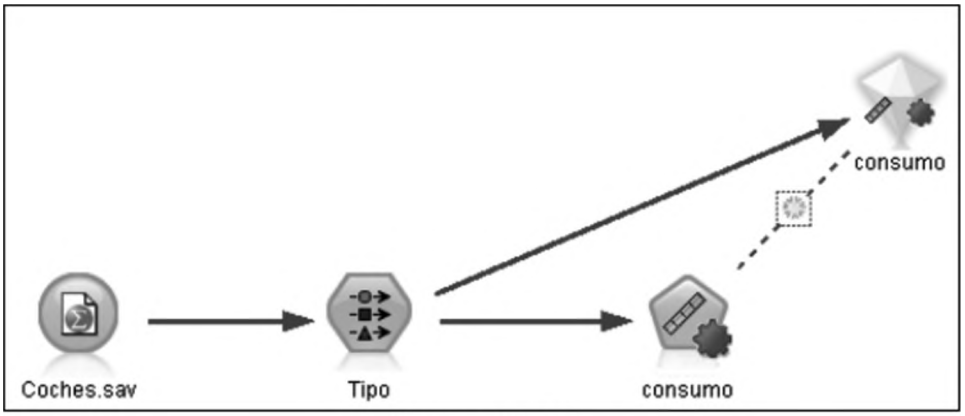


Figura 12-17

En cuanto a los **requisitos de este nodo** tenemos que solo se puede utilizar un único campo objetivo (con el papel establecido a Objetivo), y al menos un campo de entrada (con el papel establecido a Entrada). El objetivo debe ser un campo continuo (rango numérico), como edad o ingresos. Los campos de entrada pueden ser continuos o categóricos, con la limitación de que puede que algunas entradas no sean adecuadas para algunos tipos de modelo. Por ejemplo, los modelos C&RT pueden utilizar campos de cadena categóricos como entradas, mientras que los modelos Regresión lineal no pueden utilizar estos campos y los omitirán si se especifica. Los requisitos son los mismos que cuando se utilizan los nodos de modelado individuales. Por ejemplo, un modelo CHAID funciona igual independientemente de si se ha generado desde el nodo CHAID o el nodo Autonumérico.

Se pueden utilizar campos de frecuencia y ponderación. La frecuencia y la ponderación se utilizan para proporcionar importancia adicional a ciertos registros sobre otros porque, por ejemplo, el usuario sabe que el conjunto de datos creado no representa totalmente una sección de la población principal (Ponderación) o porque un registro representa un número de casos idénticos (Frecuencia). Si se especifica, los algoritmos C&RT y CHAID pueden utilizar un campo de frecuencia. Los algoritmos C&RT, CHAID, Regresión y GenLin pueden utilizar un campo de ponderación. Otros tipos de modelo omitirán estos campos y crearán los modelos de todas formas. Los campos de frecuencia y ponderación solo se utilizan para la creación de modelos y no se tienen en cuenta al evaluar o puntuar modelos. Consulte el tema Uso de campos de frecuencia y ponderación si desea obtener más información.

Los **tipos de modelo admitidos** incluyen Red neuronal, C&RT, CHAID, Regresión, GenLin, Vecino más cercano y SVM.

La pestaña *Modelo* del nodo Autonumérico (Figura 12-15) permite especificar el número de modelos que se van a guardar, junto con los criterios empleados para compararlos. Las opciones son las siguientes:

Nombre del modelo. Puede generar el nombre del modelo de forma automática basándose en el campo objetivo o de ID (o en el nombre del tipo de modelo si se especifica ningún campo objetivo), o bien especificar un nombre personalizado.

Utilizar los datos en particiones. Si se ha definido un campo de partición, esta opción garantiza que solo se utilizarán los datos de la partición de entrenamiento para la generación del modelo. Consulte el tema Nodo Partición si desea obtener más información.

Crear modelos divididos. Genera un modelo diferente para cada valor posible de campos de entrada que se especifican como modelos divididos. Consulte el tema Generación de modelos divididos si desea obtener más información.

Ordenar modelos por. Especifique los criterios utilizados para comparar modelos.

• *Correlación*. Correlación de Pearson entre el valor observado para cada registro y el valor pronosticado por el modelo. La correlación es una medida de asociación lineal entre dos variables, con valores cercanos a 1 que indican una relación más fuerte. (Los valores de correlación se encuentran entre –1, para una relación negativa perfecta, y +1, para una relación positiva perfecta. El valor 0 indica la ausencia de relaciones lineales, mientras que un modelo con una correlación negativa estaría en el último puesto de la lista).

 ©Alfaomega-RC Libros

• *Número de campos*. Número de campos utilizados como predictores en el modelo. La selección de modelos que utilizan menos campos puede simplificar la preparación de datos y mejorar el rendimiento en algunos casos.

• *Error relativo*. El error relativo es el cociente de la varianza de los valores observados de aquellos pronosticados por el modelo a la varianza de los valores observados de la media. En la práctica, compara el buen rendimiento del modelo con respecto a un modelo nulo o de intersección que simplemente devuelve el valor medio del campo objetivo como el pronóstico. En un buen modelo, este valor debe ser inferior a 1, lo que indica que el modelo es más preciso que el modelo nulo. Un modelo con un error relativo superior a 1 es menos preciso que el modelo nulo y por lo tanto no es útil. En el caso de modelos Regresión lineal, el error relativo es igual al cuadrado de la correlación y no añade información nueva. En el caso de modelos no lineales, el error relativo no está relacionado con la correlación y proporciona una medida adicional para valorar el rendimiento del modelo.

Ordenar modelos usando. Si se está usando una partición, puede especificar si los rangos se basan en la partición de entrenamiento o en la partición de comprobación. En conjuntos de datos de gran tamaño, si usa una partición para el filtrado preliminar de modelos, puede mejorar rendimiento en gran medida. Consulte el tema Nodo Partición si desea obtener más información.

Número de modelos que se utilizarán. Especifica el número máximo de modelos que aparecerán en el nugget de modelo generado por el nodo. Los primeros modelos de la lista se enumeran en función del criterio de ordenación especificado. El aumento de este límite le permitirá comparar resultados de más modelos pero puede ralentizar el rendimiento. El valor máximo permitido es 100.

Calcular importancia del predictor. En el caso de modelos que produzcan una medida adecuada de importancia, puede mostrar un gráfico que indique la importancia relativa de cada predictor al estimar el modelo. Normalmente, desea centrar sus esfuerzos de modelado en los predictores que importan más y considera eliminar o ignorar los que importan menos. Tenga en cuenta que la importancia de predictor puede aumentar el tiempo necesario para calcular algunos modelos; además no se recomienda si solo desea una amplia comparación entre varios modelos diferentes. Es de mayor utilidad una vez ha limitado su análisis a unos cuantos modelos que desee explorar más a fondo. Consulte el tema Importancia del predictor si desea obtener más información.

No conservar modelos si. Especifica valores de umbral para la correlación, el error relativo y el número de campos utilizados. Los modelos que no cumplen alguno de estos criterios se descartarán y no se incluirán en el informe de resumen.

• *Correlación menor que*. Correlación mínima (en cuanto a valor absoluto) para que un modelo se incluya en el informe de resumen.

• *Número de campos utilizados mayor que*. Número máximo de campos que puede utilizar cualquier modelo que vaya a incluirse.

• *Error relativo mayor que*. Error relativo máximo para cualquier modelo que vaya a incluirse.

Si lo desea, puede configurar el nodo para que se detenga la ejecución la primera vez que se genere un modelo que cumpla todos los criterios especificados.

La pestaña *Experto* del nodo Autonumérico (Figura 12-16) le permite seleccionar los algoritmos y opciones que se van a usar y especificar las reglas de parada. Las opciones son las siguientes:

Modelos utilizados. Use las casillas de verificación de la columna izquierda para seleccionar los tipos de modelo (algoritmos) que se van a incluir en la comparación. Cuantos más tipos seleccione, más modelos se crearán y más tardará el procesamiento.

Tipo de modelo. Enumera los algoritmos disponibles (consulte a continuación).

Parámetros del modelo. Puede usar la configuración por defecto o seleccionar Especificar para elegir opciones para cada tipo de modelo. Las opciones específicas son parecidas a las disponibles en los nodos de modelado independientes, con la diferencia de que se pueden seleccionar varias opciones o combinaciones. Por ejemplo, si compara los modelos del nodo Red neuronal, puede seleccionar los seis modelos para entrenarlos de una vez en lugar de seleccionar uno de ellos.

Número de modelos. Enumera el número de modelos generados para cada algoritmo basados en la configuración actual. Al combinar opciones, puede aumentar rápidamente el número de modelos, por lo que se recomienda prestar especial atención a este número, especialmente si usa conjuntos de datos grandes.

 ©Alfaomega-RC Libros

Limitar el tiempo máximo empleado en generar un único modelo. (Solo modelos de K-medias, Kohonen, bietápicos, SVM, KNN, de red bayesiana y de lista de decisiones) Establece un límite de tiempo máximo para cualquier modelo. Por ejemplo, si un modelo determinado necesita un período de tiempo más largo del esperado para entrenarse debido a una interacción compleja, es probable que no quiera detener la ejecución de todo el modelado.

Se admiten los siguientes algoritmos:

Red neuronal. El nodo Red neuronal utiliza un modelo simplificado que emula el modo en que el cerebro humano procesa la información. Funciona simultaneando un número elevado de unidades simples de procesamiento interconectadas que parecen versiones abstractas de neuronas. Las redes neuronales son estimadores potentes de funciones generales y requieren un conocimiento matemático o estadístico mínimo para entrenarlas o aplicarlas.

Nodo C&RT. El nodo de árbol de clasificación y regresión (C&R) genera un árbol de decisión que permite pronosticar o clasificar observaciones futuras. El método utiliza la partición reiterada para dividir los registros de entrenamiento en segmentos minimizando las impurezas en cada paso, donde un nodo se considera "puro" si el 100% de los casos del nodo corresponden a una categoría específica del campo objetivo. Los campos de entrada y objetivo pueden ser continuos (rango numérico) o categóricos (nominal, ordinal o marca). Todas las divisiones son binarias (solo se crean dos subgrupos).

Nodo CHAID. El nodo CHAID genera árboles de decisión utilizando estadísticos de chi-cuadrado para identificar las divisiones óptimas. A diferencia de los nodos C&RT y QUEST, CHAID puede generar árboles no binarios, lo que significa que algunas divisiones generarán más de dos ramas. Los campos de entrada y objetivo pueden ser continuos (rango numérico) o categóricos. CHAID exhaustivo es una modificación de CHAID que examina con mayor precisión todas las divisiones posibles, aunque necesita más tiempo para realizar los cálculos.

Nodo Regresión lineal. La regresión lineal es una técnica de estadístico común utilizada para resumir datos y realizar pronósticos ajustando una superficie o línea recta que minimice las discrepancias existentes entre los valores de salida reales y los pronosticados. Consulte el tema Nodo Regresión si desea obtener más información.

Modelo lineal generalizado. El modelo lineal generalizado amplía el modelo lineal general, de manera que la variable dependiente está relacionada linealmente con los factores y las covariables mediante una determinada función de enlace. Además, el modelo permite que la variable dependiente tenga una distribución que no sea normal. Cubre la funcionalidad de un amplio número de modelos estadísticos, incluyendo regresión lineal, regresión logística, modelos log lineales para recuento de datos y modelos de supervivencia censurados por intervalos. Consulte el tema Nodo GenLin si desea obtener más información.

Nodo KNN. El nodo k de modelado de vecino (KNN) asocia el nuevo caso con la categoría o valor de los objetos k junto a él en el espacio de predictores, donde k es un entero. Los casos parecidos están próximos y los que no lo son están alejados entre sí. Consulte el tema Nodo KNN si desea obtener más información.

Nodo SVM. El nodo Máquina de vectores de soporte (SVM) le permite clasificar datos en uno o dos grupos sin que haya un ajuste por exceso. SVM funciona bien con conjuntos de datos grandes, como aquellos con un gran número de campos de entrada. Consulte el tema Nodo SVM si desea obtener más información.

Modelos lineales. Los modelos de regresión lineal predicen un destino continuo tomando como base las relaciones lineales entre el destino y uno o más predictores. Consulte el tema Modelos lineales si desea obtener más información.

Si ahora hacemos clic con el botón derecho del ratón en el nodo *Autonumérico* sobre la ruta, se obtiene la pantalla de resultados de la Figura 12-18.

En la solapa *Modelo* de esta figura se observa que el mejor modelo predictivo estimado es la Red neuronal, seguido del Modelo lineal generalizado y del modelo de Regresión. En las diferentes columnas de la Figura 12-18 se observan los parámetros de evaluación los modelos seleccionados.

En la solapa *Gráfico* (Figura 12-19) se observan los gráficos de dispersión de la variable predicha respecto de sus valores reales (ha de ajustarse bien a la diagonal del primer cuadrante) y de importancia del predictor.

©Alfaomega-RC Libros

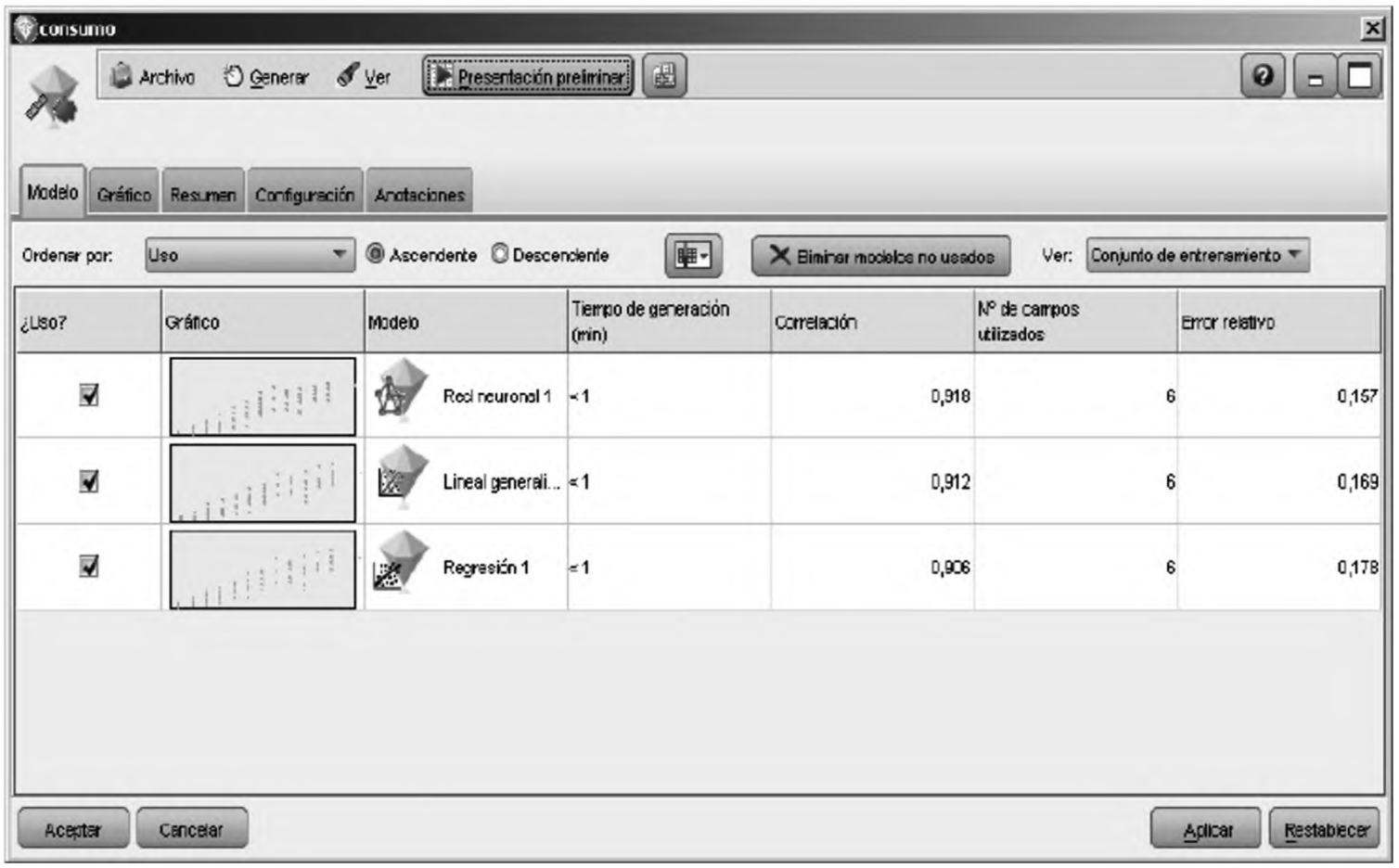


Figura 12-18

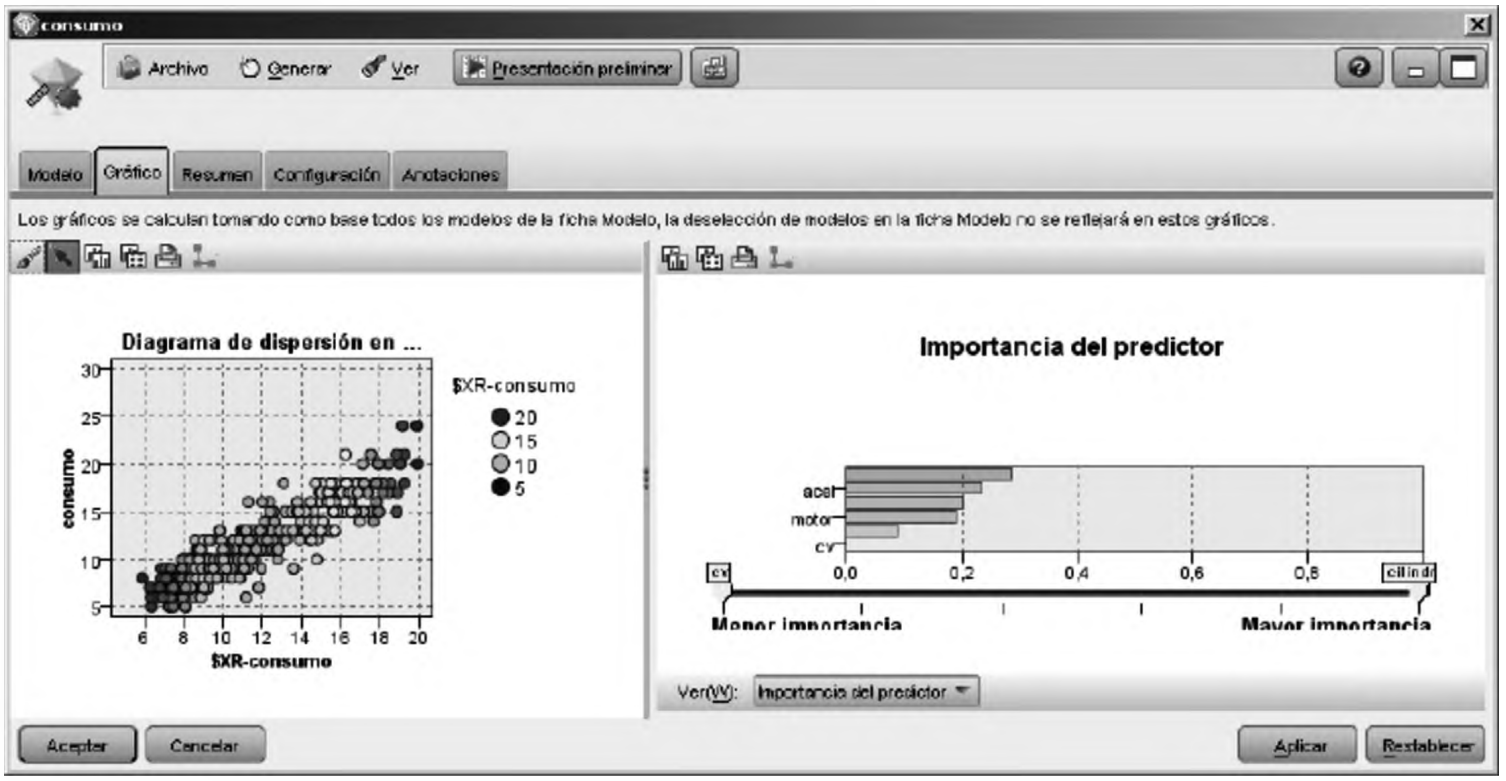


Figura 12-19

Si se añade un nodo *Tabla* al nodo *Autonumérico* (Figura 12-20) se observan dos nuevos campos que contienen los valores predichos por el modelo para la variable dependiente y el error del modelo.

Tabla (11 campos, 406 registros)

Archivo Edición Generar

Tabla Anotaciones

	consu...	motor	cv	peso	acel	año	origen	cilindr	derivada	$XR-consumo	$XRE-consumo
1	13	5031	130	1169	12	70	1	8	0	13.745	0.141
2	16	5735	165	1231	12	70	1	8	0	15.751	0.084
3	13	5211	150	1145	11	70	1	8	0	14.530	0.019
4	15	4982	150	1144	12	70	1	8	0	14.634	0.032
5	14	4949	140	1149	10	70	1	8	0	13.998	0.062
6	16	7030	198	1447	10	70	1	8	0	17.830	0.267
7	17	7440	220	1451	9	70	1	8	0	18.449	0.507
8	17	7210	215	1437	8	70	1	8	0	18.140	0.475
9	17	7456	225	1475	10	70	1	8	0	18.914	0.482
10	16	6391	190	1283	8	70	1	8	0	16.544	0.173

Figura 12-20

EL ALGORITMO AUTOCONGLOMERACIÓN DE IBM SPSS MODELER

El nodo Autoconglomeración calcula y compara los modelos de conglomerado que identifican grupos de registros con características similares. El nodo funciona de la misma manera que otros nodos de modelado de conglomerado, permitiéndole experimentar con múltiples combinaciones de opciones en una única pasada de modelado. Los modelos se pueden comparar utilizando medidas básicas con las que se intenta filtrar y definir la utilidad de los modelos de conglomerado y proporcionar una medida según la importancia de campos concretos.

Los modelos de conglomerado se suelen identificar con grupos que se pueden utilizar como entradas en futuros análisis. Por ejemplo, es posible que desee dirigirse a grupos de clientes según sus características demográficas como ingresos, o según los servicios que hayan contratado en el pasado. Esto puede hacerse si se tiene un conocimiento previo de los grupos y sus características; es posible que no sepa en cuántos grupos buscar o las funciones que debe utilizar para definirlos. Los modelos de conglomerado se suelen definir como modelos de aprendizaje no supervisado, ya que no utilizan un campo de destino y no devuelven una predicción específica que se pueda evaluar como true o false. El valor de un modelo de conglomerado viene determinado por su capacidad de capturar agrupaciones interesantes en los datos y proporcionar descripciones útiles de dichas agrupaciones. Consulte el tema Modelos de conglomerados si desea obtener más información.

Como ejemplo consideramos el archivo en formato SAS *mundo.sas7bdat* que contiene información sobre variables que inciden en el nivel de desarrollo de los países del mundo. Trataremos de buscar conglomerados que agrupen los países por nivel similar de desarrollo según las variables densidad de población (*densidad*), nivel de

 ©Alfaomega-RC Libros

población urbana (*urbana*), esperanza de vida femenina (*espvidif*), esperanza de vida masculina (*espvidam*), nivel de alfabetización (*alfabet*), tasa de mortalidad infantil (*mortinf*), tasa de natalidad (*tasa_nat*), tasa de mortalidad (*tasa_mort*), tasa de sida (*tasasida*) y producto interior bruto en logaritmos (*log_pib*).

Una vez importado el fichero como un origen de datos SAS, podemos utilizar el nodo *Tipo* para asignar a la variable *PAIS* el papel de variable *objetivo* y al resto de las variables de clasificación el papel de *entrada*. A continuación se une al diagrama el nodo *Clasificador automático*, se hace doble clic sobre él con el botón secundario del ratón y se elige *Edición* en el menú emergente resultante para definir los parámetros en la Figura 12-21. En la solapa *Campos* elegimos *Utilizar configuración del nodo Tipo* si queremos utilizar las variables definidas en el nodo *Tipo* o elegimos la opción *Utilizar configuración personalizada* definiendo la variable de *Evaluación* (objetivo) y las variables de *Entrada* (variables de clasificación) tal y como se indica en la Figura 12-21. En la solapa *Modelo* (Figura 12-22) definimos las características generales de los modelos a ajustar. En la solapa *Experto* (Figura 12-23) se seleccionan las opciones para cada tipo de modelo, en la solapa *Descartar* (Figura 12-24) se definen condiciones de exclusión de modelos en el análisis. Al hacer clic en *Aplicar* y *Ejecutar* se obtienen las pantallas temporales que informan del proceso de ejecución. Finalmente se obtiene el modelo sobre la ruta (Figura 12-25).

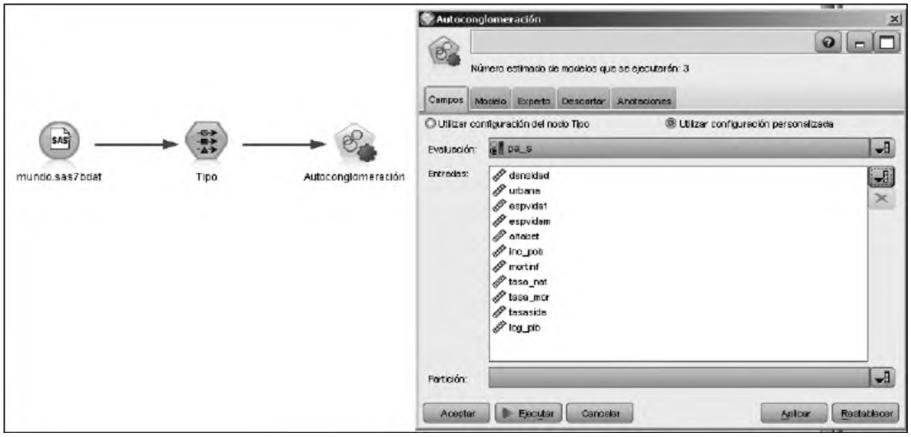


Figura 12-21

Figura 12-22

Figura 12-23

Figura 12-24

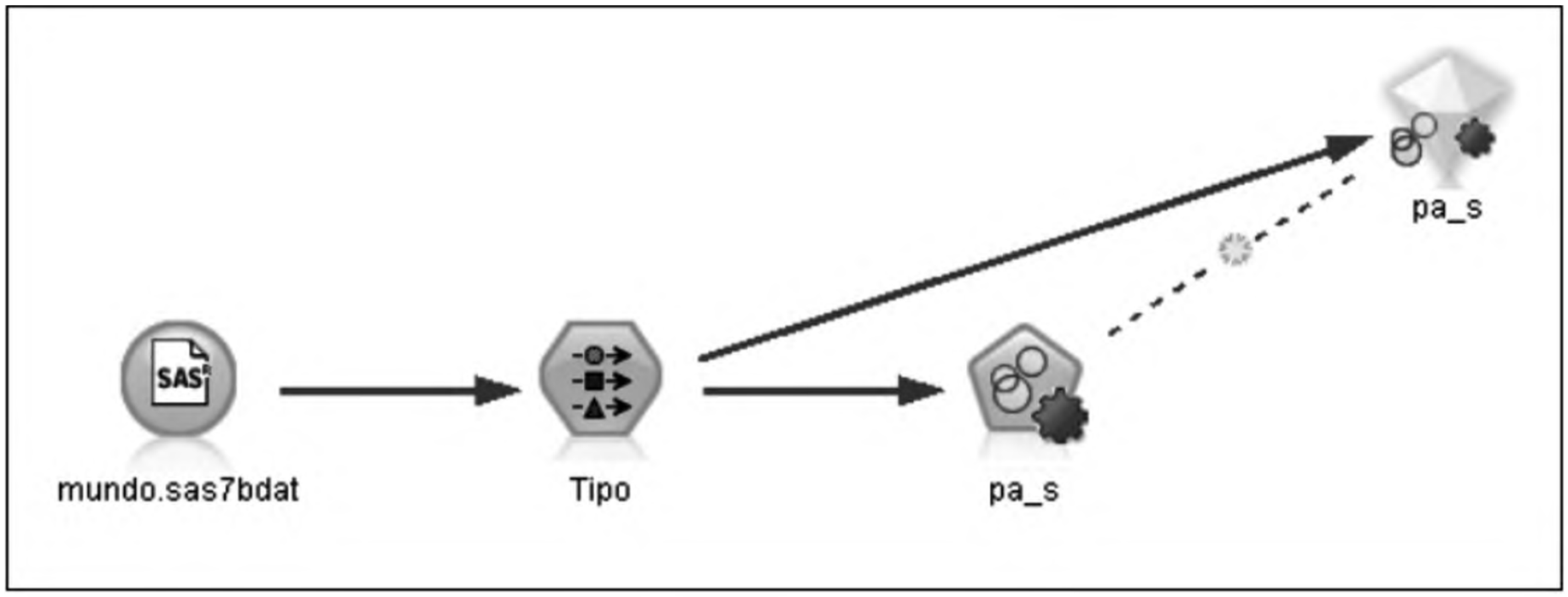


Figura 12-25

©Alfaomega-RC Libros

Los modelos de conglomerados no utilizan campos objetivo de la misma manera que otros modelos, porque no realizan predicciones específicas que se pueden evaluar como true o false. En su lugar, se utilizan para identificar grupos de casos que pueden estar relacionados. Por ejemplo, no puede utilizar un modelo de conglomerado para predecir si un cliente concreto abandonará o responderá a una oferta. Pero puede utilizar un modelo de conglomerado para asignar clientes a grupos en función de su tendencia a hacer determinadas cosas. Los campos de ponderación y frecuencia no se usan.

En cuanto a los *Campos de evaluación*, mientras no utilice un objetivo, puede especificar uno o más campos de evaluación que se utilizarán en la comparación de modelos. La utilidad de un modelo de conglomerado se puede evaluar lo bien (o mal) que los conglomerados diferencian los campos.

Entre los tipos de modelos admitidos se incluyen TwoStep, K-Means y Kohonen.

La pestaña *Modelo* del nodo Autoconglomeración binario (Figura 12-22) le permite especificar el número de modelos que se van a guardar, junto con los criterios empleados para compararlos. Sus campos son los siguientes:

Nombre del modelo. Puede generar el nombre del modelo de forma automática basándose en el campo objetivo o de ID (o en el nombre del tipo de modelo si se especifica ningún campo objetivo), o bien especificar un nombre personalizado.

Utilizar los datos en particiones. Si se ha definido un campo de partición, esta opción garantiza que solo se utilizarán los datos de la partición de entrenamiento para la generación del modelo. Consulte el tema Nodo Partición si desea obtener más información.

Ordenar modelos por. Especifique los criterios utilizados para comparar y clasificar los modelos. Se admiten las siguientes entradas:

• *Silueta*. Un índice que mide la cohesión y separación del conglomerado. Consulte Medida de ordenación de siluetas a continuación para obtener más información.

• *Número de conglomerados*. El número de conglomerados que se utilizan en el modelo.

• *Tamaño del conglomerado más pequeño*. El menor tamaño de conglomerado.

• *Tamaño del conglomerado mayor*. El mayor tamaño de conglomerado.

• *Conglomerado mayor / menor*. La razón del tamaño del conglomerado menor y el mayor.

• *Importancia*. La importancia del campo *Evaluación* en la pestaña *Campos*. Tenga en cuenta que solo se puede calcular si se ha especificado un campo *Evaluación*.

Ordenar modelos usando. Si se está usando una partición, puede especificar si los rangos se basan en el conjunto de datos de entrenamiento o en el conjunto de prueba. En conjuntos de datos de gran tamaño, si usa una partición para el filtrado preliminar de modelos, puede mejorar rendimiento en gran medida.

Número de modelos que se mantendrán. Especifica el número máximo de modelos que aparecerán en el nugget generado por el nodo. Los primeros modelos de la lista se enumeran en función del criterio de ordenación especificado. Tenga en cuenta que si aumenta este límite puede ralentizarse el rendimiento. El valor máximo permitido es 100.

La pestaña *Experto* del nodo Autoconglomeración le permite aplicar una partición (si está disponible), seleccionar los algoritmos que se va a usar y especificar las reglas de parada. Sus campos son los siguientes:

Modelos utilizados. Use las casillas de verificación de la columna izquierda para seleccionar los tipos de modelo (algoritmos) que se van a incluir en la comparación. Cuantos más tipos seleccione, más modelos se crearán y más tardará el procesamiento.

Tipo de modelo. Enumera los algoritmos disponibles.

Parámetros del modelo. Puede usar la configuración por defecto o seleccionar *Especificar* para elegir opciones para cada tipo de modelo. Las opciones específicas son parecidas a las disponibles en los nodos de modelado independientes, con la diferencia de que se pueden seleccionar varias opciones o combinaciones. Por ejemplo, si compara los modelos del nodo Red neuronal, puede seleccionar los seis modelos para entrenarlos de una vez en lugar de seleccionar uno de ellos.

Número de modelos. Enumera el número de modelos generados para cada algoritmo basados en la configuración actual. Al combinar opciones, puede aumentar

 ©Alfaomega-RC Libros

rápidamente el número de modelos, por lo que se recomienda prestar especial atención a este número, especialmente si usa conjuntos de datos grandes.

Limitar el tiempo máximo empleado en generar un único modelo. (Solo modelos de K-medias, Kohonen, bietápicos, SVM, KNN, de red bayesiana y de lista de decisiones) Establece un límite de tiempo máximo para cualquier modelo. Por ejemplo, si un modelo determinado necesita un período de tiempo más largo del esperado para entrenarse debido a una interacción compleja, es probable que no quiera detener la ejecución de todo el modelado.

Entre los algoritmos admitidos tenemos:

Nodo K-medias. El nodo K-medias agrupa conjuntos de datos en grupos distintos (o conglomerados). El método define un número fijo de conglomerados, de forma iterativa asigna registros a los conglomerados y ajusta los centros de los conglomerados hasta que no se pueda mejorar el modelo. En lugar de intentar pronosticar un resultado, los modelos de k-medias utilizan un proceso conocido como aprendizaje no supervisado para revelar los patrones del conjunto de campos de entrada.

Nodo Kohonen. El nodo Kohonen genera un tipo de red neuronal que se puede usar para conglomerar un conjunto de datos en grupos distintos. Cuando la red se termina de entrenar, los registros que son similares se deberían cerrar juntos en el mapa de resultados, mientras que los registros que son diferentes aparecerían aparte. Puede observar el número de observaciones capturadas por cada unidad en el nugget de modelo para identificar unidades fuertes. Esto le proporcionará una idea del número apropiado de conglomerados.

Nodo Bietápico. El nodo Bietápico es un método de conglomerado de dos pasos. El primer paso es hacer una única pasada por los datos para comprimir los datos de entrada de la fila en un conjunto de subconglomerados administrable. El segundo paso utiliza un método de conglomerado jerárquico para fundir progresivamente los subconglomerados en conglomerados cada vez más grandes. El bietápico tiene la ventaja de estimar automáticamente el número óptimo de conglomerados para los datos de entrenamiento. Puede gestionar tipos de campos mixtos y grandes conjuntos de datos eficazmente.

La pestaña *Descartar* del nodo Autoconglomeración le permite descartar automáticamente los modelos que no cumplen determinados criterios. Puede especificar el valor mínimo de silueta, número de conglomerados, tamaños de conglomerados y la importancia del campo de evaluación en el modelo. La silueta y el número y tamaño de clusters están determinados en la especificación del nodo de

modelado. Si lo desea, puede configurar el nodo para que se detenga la ejecución la primera vez que se genere un modelo que cumpla todos los criterios especificados.

Si ahora hacemos clic con el botón derecho del ratón en el nodo *Autoconglomeración* sobre la ruta, se obtiene la pantalla de resultados de la Figura 12-26. En la solapa *Modelo* de esta figura se observa que el mejor modelo de conglomeración es el método *Bietápico*, seguido del algoritmo K-medias y de la red neuronal Kohonen. En la solapa *Gráfico* se observan los histogramas de frecuencias cuyo número de barras indican el número de conglomerados óptimo para cada método y cuya longitud de las barras indican el porcentaje de países clasificados en cada conglomerado. La última columna de la Figura 12-26 indica el índice de importancia del método y se utiliza para ordenar los distintos métodos de conglomeración según su eficacia.

Si se añade un nodo *Tabla* al nodo *Autoconglomeración* (Figura 12-27) se observa un nuevo campo al final de la tabla de datos que contienen la pertenencia de cada registro (país) al conglomerado correspondiente.

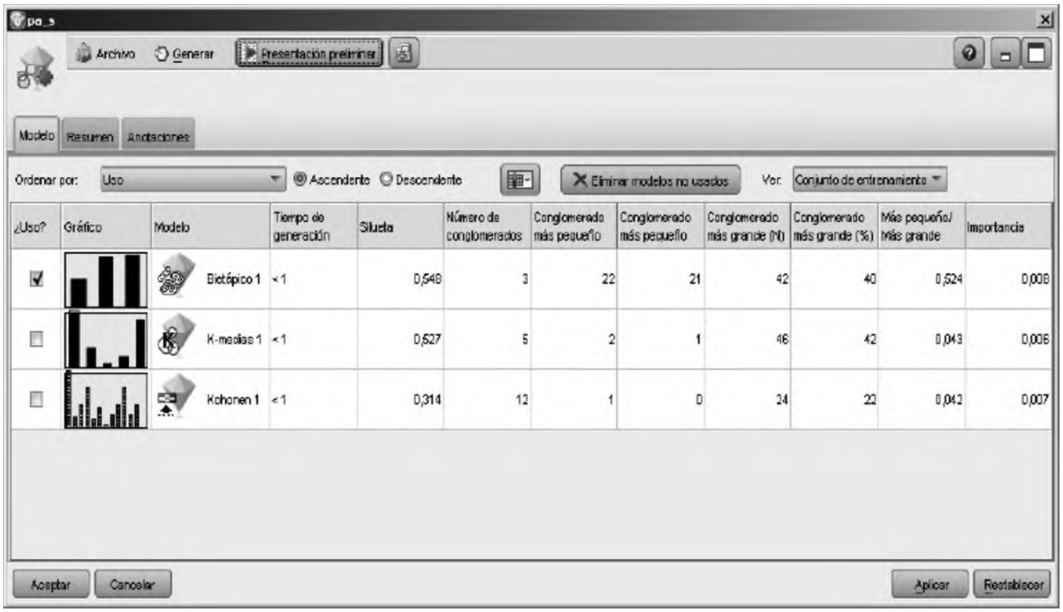


Figura 12-26

©Alfaomega-RC Libros

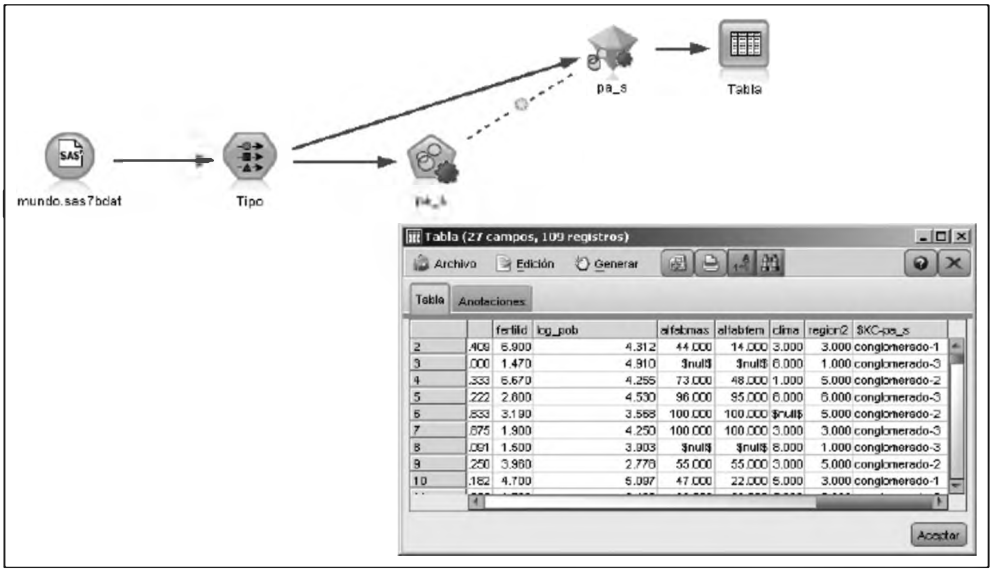


Figura 12-27

Cuando se ejecuta el nodo de modelado automático de conglomeración, el nodo estima modelos de candidato de cada combinación de opciones posible, clasifica cada modelo de candidato en función de la medida que especifique y guarda los mejores modelos (Figura 12-26). La solapa *Modelo* de la salida contiene un conjunto de uno o más modelos que genera el nodo, que se pueden examinar o seleccionar individualmente para la puntuación. El tipo de modelo y el tiempo de generación se incluyen para cada modelo, junto con un número de otras mediciones que resulten adecuadas para el modelo. Puede ordenar la tabla en cualquiera de estas columnas para identificar rápidamente los modelos más interesantes. Se tendrá en cuenta lo siguiente:

• Para examinar cualquiera de los nugget de modelo individuales, pulse dos veces en el icono del nugget. A partir de aquí, puede generar un nodo de modelado para ese modelo en el lienzo de rutas, o una copia del nugget de modelo en la paleta de modelos.

• Los gráficos de miniatura ofrecen una rápida valoración visual de cada modelo, tal y como se resume a continuación. Puede pulsar dos veces en una miniatura para generar un gráfico a tamaño completo. El gráfico a tamaño completo muestra hasta 1000 puntos y se basará en una muestra si el conjunto de datos contiene más. Solo en el caso de los diagramas de dispersión, el gráfico se regenera cada vez que se muestra, de modo que cualquier cambio en los datos anteriores en la ruta (como la actualización de una muestra aleatoria o partición si Establecer semilla aleatoria no está seleccionado) pueda reflejarse cada vez que se vuelva a dibujar el diagrama de dispersión.

• Use la barra de herramientas para mostrar u ocultar columnas específicas de la pestaña *Modelo* o cambiar la columna usada para ordenar la tabla. (También puede cambiar el orden pulsando en las cabeceras de columna).

• Utilice el botón *Eliminar* para eliminar permanentemente los modelos no utilizados.

• Para reorganizar columnas, pulse en la cabecera de una columna y arrastre la columna a la ubicación que desee.

• Si se está usando una partición, puede optar por ver los resultados de la partición de comprobación o entrenamiento según proceda. Consulte el tema Nodo Partición si desea obtener más información.

Las columnas específicas de la Figura 12-26 dependen del tipo de modelos que se estén comparando y de la naturaleza de los objetivos, tal y como se indica a continuación.

Objetivos binarios

• En el caso de modelos binarios, la miniatura muestra la distribución de valores reales, superpuestos con los valores pronosticados, para ofrecer una rápida indicación visual de cuántos registros se pronosticaron correctamente en cada categoría.

• Los criterios de clasificación coinciden con las opciones del nodo de modelado Clasificador automático.

• Para obtener el máximo beneficio, también aparece en el informe el percentil en el que se produce el valor máximo.

• En el caso de la elevación acumulada, puede cambiar el percentil seleccionado mediante la barra de herramientas.

Objetivos nominales

• En el caso de modelos nominales (conjunto), la miniatura muestra la distribución de valores reales, superpuestos con los valores pronosticados, para ofrecer una rápida indicación visual de cuántos registros se pronosticaron correctamente en cada categoría.

 ©Alfaomega-RC Libros

• Los criterios de clasificación coinciden con las opciones del nodo de modelado Clasificador automático.

Objetivos continuos

• En el caso de modelos continuos (rango numérico), el gráfico representa los valores pronosticados frente a los valores observados de cada modelo, lo que ofrece una rápida indicación visual de la correlación entre ellos. En el caso de un buen modelo, los puntos tienden a conglomerarse en la diagonal en lugar de estar dispersos aleatoriamente por el gráfico.

• Los criterios de clasificación coinciden con las opciones del nodo de modelado Autonumérico.

Objetivos de conglomerado

• En el caso de modelos de conglomerado, el gráfico representa los recuentos frente a los conglomerados para cada modelo, lo que ofrece una rápida indicación visual de la distribución de conglomerados.

• Los criterios de clasificación coinciden con las opciones del nodo de autoconglomerado.

Selección de modelos para puntuación

La columna Uso? le permite seleccionar los modelos que se utilizarán en la puntuación.

• En objetivos numéricos, nominales y binarios, podrá seleccionar múltiples modelos de puntuación y combinar los resultados en el nugget de modelo de conjunto único. Al combinar predicciones de varios modelos, pueden evitarse las limitaciones en modelos individuales que suelen dar como resultado una precisión global superior que puede obtenerse de cualquiera de los modelos.

• En modelos de conglomerado solo puede seleccionar un modelo de puntuación cada vez. Por defecto, el primer clasificado se selecciona primero.

FASE DE EVALUACIÓN EN SAS ENTERPRISE MINER. NODO ASSESSMENT

La evaluación del modelo y su comparación con otros es una de las etapas más importantes de la Minería de Datos de cara a la posterior toma de decisiones. Este nodo permite comparar modelos creados en nodos diferentes (Figura 12-28).

El nodo *Assesment* utiliza para realizar la valoración la definición que del output se haya hecho en un nodo anterior. El nodo no permite por tanto redefinir el objetivo (salvo en alguna excepción a la hora de realizar gráficos).

El nodo de evaluación requiere un fichero de datos de predicción. Este fichero se genera cuando entrenamos un modelo. La valoración nos mostrará una tabla de los resultados de ajuste de los modelos y distintos gráficos dependiendo de si la variable es o no de intervalo y de si hemos definido una matriz de pérdidas o beneficios en la opción *Target Profile*.

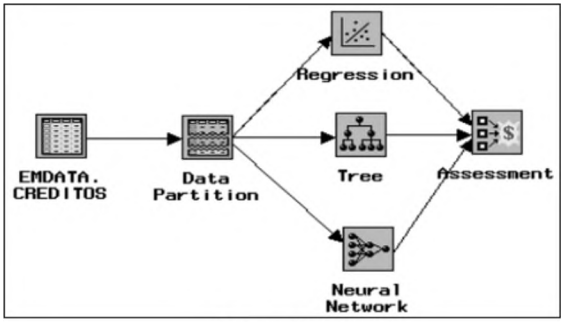


Figura 12-28

TRABAJO CON EL NODO DE EVALUACIÓN

El fichero de datos con las predicciones

El fichero *Score* contiene las probabilidades a posteriori de las variables de resultado binarias, ordinales y nominales. La suma de estas probabilidades siempre será la unidad. Los nodos predictivos Regression, Tree, Neural Network, Ensemble y User-Defined Model automáticamente producen este fichero como resultado. Si la variable objetivo es una variable continua en lugar de probabilidades se calculará el valor de la predicción.

©Alfaomega-RC Libros

Una vez abierto el nodo de valoración veremos que este consta de cuatro pestañas:

Models

Muestra el nombre del modelo y la herramienta predictiva que utiliza junto a un número de estadísticas de valoración (Figura 12-29). El nodo *Ensemble* no envía estas estadísticas al nodo de valoración.

Figura 12-29

Por defecto esta pestaña proporciona la siguiente información:

TOOL: nombre de la herramienta de predicción.

NAME: nombre del modelo que puede ser cambiado en el nodo correspondiente.

DESC: descripción del modelo.

TARGET1: nombre de la variable objetivo.

_TARGET1_LEVEL1_: categoría de referencia.

_BY_GROUP_ID_: etiqueta identificativa de grupo creada en el nodo Group Processing.

_BY_GROUP_DESC: descripción de cada uno de los grupos creados en el nodo Group Processing.

RASE (Square root of average squared error): raíz cuadrada de la media de errores al cuadrado de los datos de entrenamiento.

$$RASE = \sqrt{ASE} = \sqrt{SSE / n}$$

VRASE: raíz cuadrada de la media de errores al cuadrado de los datos de validación.

TRASE: raíz cuadrada de la media de errores al cuadrado de los datos de test.

SBC: criterio bayesiano de Schwarz $SBC = n\ln(SSE/n) + k\ln(n)$.

MISC: proporción de datos de entrenamiento clasificados de forma incorrecta.

VMISC: proporción de datos de validación clasificados de forma incorrecta.

TMISC : proporción de datos de test clasificados de forma incorrecta.

_N_CASES_TRAIN_: número de casos de entrenamiento.

_N_CASES_VALIDATE_: número de casos de validación.

_N_CASES_TEST_: número de casos test.
Utilizando la opción del menú principal *View* $\rightarrow$ *Show All* tenemos:

AIC: criterio de información de Akaike. $AIC = n\ln(SSE/n) + 2k$.

AVERR: average error function.

DFE: grados de libertad del término de error. $n - k$.

DFM: grados de libertad del modelo. K.

DFT: grados de libertad totales.

DIV: divisor para calcular el promedio del error al cuadrado (observaciones).

ERR: función del error. $ERR = \frac{2}{\pi}\int_0^x e^{-z^2}dz$

FPE: error de predicción final. $FPE = \frac{SSE(n+k)}{n(n-k)}$

MAX: error absoluto máximo. $MAX = \max|y_i - \hat{y}_i|$

MSE: error cuadrado medio. $MSE = \frac{SSE}{n-k}$

NOBS: suma de frecuencias.

_NW_: número de pesos estimados.

RFPE: raíz cuadrada del error de predicción final. $RFPE = \sqrt{FPE}$

©Alfaomega-RC Libros

RMSE: raíz cuadrada del error al cuadrado medio. $RMSE = \sqrt{MSE}$

SSE: suma de errores al cuadrado.

SUMW: suma de la frecuencia de casos ponderados.

WRONG: número de casos mal clasificados.

Donde *n* es el número de observaciones y *k* el número de parámetros. Para poder ver los resultados gráficos de un modelo, deberemos seleccionarlo. Para elegir un modelo haremos clic con el botón izquierdo del ratón y con la tecla *Control* pulsada seleccionaremos los modelos deseados. Si se trata de modelos adyacentes podremos elegir los modelos deseados arrastrando la selección. Al seleccionar más de un modelo las pestañas *Options y Reports* quedarán desactivadas.

Para ver los resultados tal y como aparecen en el visor de resultados de cada modelo seleccionaremos en el menú principal *View* →*Model Results*.

Para exportar tanto los resultados como los datos asociados, una vez seleccionado el modelo en el menú principal iremos a *File* → *Export Model* (Figura 12-30).

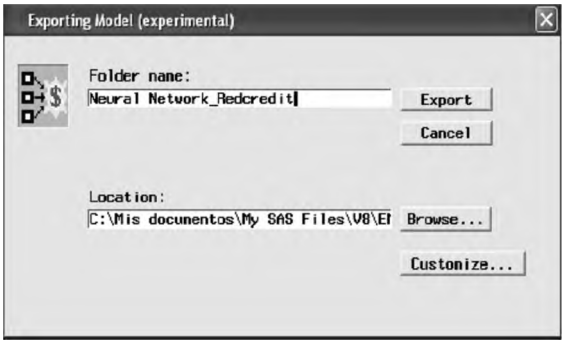


Figura 12-30

Una vez escrito el nombre del archivo y su localización podemos abrir la ventana *Customize* para seleccionar la información que queremos guardar (Figura 12-31).

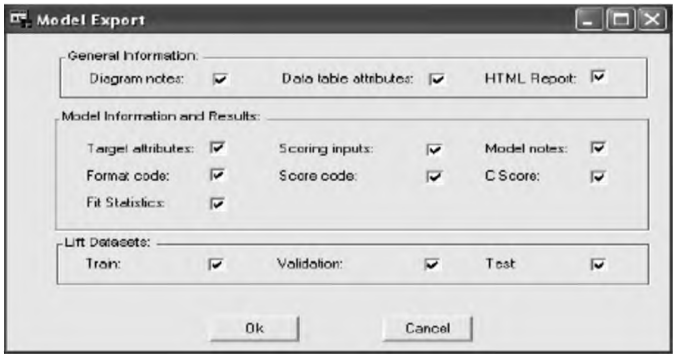


Figura 12-31

Esta utilidad creará una carpeta con tres ficheros: un informe con la información HTML del nodo, un fichero SAS con las estadísticas, predicciones y macros y finalmente un fichero de texto con las estadísticas de ajuste, código de predicción en C e información sobre el diagrama, modelo y variables utilizadas.

Options

Indica los datos (entrenamiento, validación o test) utilizados para construir los gráficos. Estas opciones no pueden ser cambiadas en el nodo de valoración. Para poder variar estas opciones es necesario acudir al Model Manager del nodo que contiene el modelo.

Reports

Muestra los gráficos y estadísticas disponibles para el modelo seleccionado. Para poder variar estas opciones es necesario acudir al Model Manager del nodo que contiene el modelo.

Output

Selecciona el modelo que queremos que pase a nodos sucesivos como el nodo *Store* (Figura 12-32).

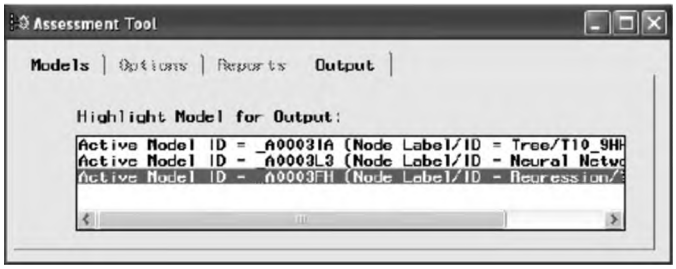


Figura 12-32

©Alfaomega-RC Libros

GRÁFICOS DEL NODO ASSESSMENT

El tipo de gráfico puede seleccionarse en el menú *Tools*. En este menú también podemos especificar la partición de los datos (*Validación* es la opción por defecto). Las selecciones deben activarse en el *Model Manager* de cada modelo o en las opciones por defecto del nodo *Assessment*. Una vez cambiadas las opciones el nodo *Assessment* deber ser ejecutado de nuevo. Podemos dividir los gráficos que el nodo *Assessment* elabora en los siguientes tipos:

LIFT (Gráfico de ganancias)

En los gráficos de este tipo las predicciones son ordenadas atendiendo a la probabilidad a posteriori calculada de la respuesta de referencia. Una vez ordenadas las observaciones son agrupadas en decilas. Un gráfico de ganancias cruzado (cross lift chart) sería lo mismo pero en esta ocasión se compararían los datos de entrenamiento y validación de un mismo modelo. Por defecto estos gráficos son de dos dimensiones aunque también pueden trazarse en tres dimensiones *Format → 3D. Enterprise Miner* puede dibujar este gráfico utilizando en ordenadas distintas medidas:

- ***%Response***: indica que en el nivel de las decilas señaladas en abscisas el modelo predice adecuadamente el porcentaje mostrado en ordenadas (Figura 12-33).

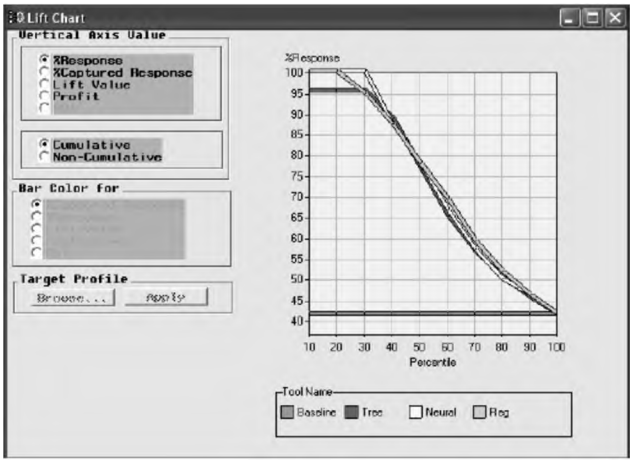


Figura 12-33

En el gráfico anterior comprobamos cómo el modelo de redes neuronales (*Neural*) predice de forma correcta el 100% de los casos con la categoría de referencia hasta el percentil 30. La línea *Baseline* indica cuál sería la capacidad de acierto de un modelo aleatorio el cual coincide con la probabilidad de pertenencia a priori. Esta referencia puede ser eliminada si no deseamos que aparezca en el menú principal *Format* →*Show Baseline*.

El gráfico anterior también puede ser mostrado de forma no acumulada (*Non-Cumulative*), es decir, mostrando el porcentaje de respuestas pertenecientes a la categoría de referencia bien clasificado en cada decila. El gráfico también muestra interactivamente el valor de cada punto del gráfico. Para ello seleccionaremos el icono *View Info* de la barra de herramientas y seleccionaremos, manteniendo pulsado el botón izquierdo, un punto del gráfico. El gráfico muestra el valor de los puntos si nos desplazamos a lo largo del mismo. El gráfico también permite variar, para un solo modelo a la vez, la categoría de referencia del output seleccionando *Edit* en la opción *Target Profile*. Sin embargo, para cambiarlo definitivamente debemos redefinirlo en el Model Manager o en el nodo Input Data Source. En esta ocasión mostramos también la línea de referencia exacta seleccionando en el menú principal *Format* →*Show Exact*. Es decir, la forma que tendría un modelo capaz de predecir correctamente a todos los individuos (Figura 12-34).

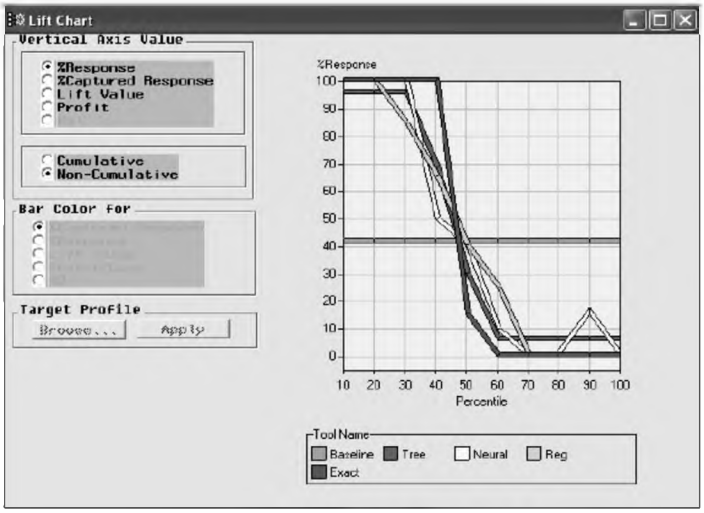


Figura 12-34

©Alfaomega-RC Libros

- ***%Captured Response***: en esta ocasión en ordenadas se representa el porcentaje de casos pertenecientes al total de la categoría de referencia (Figura 12-35).

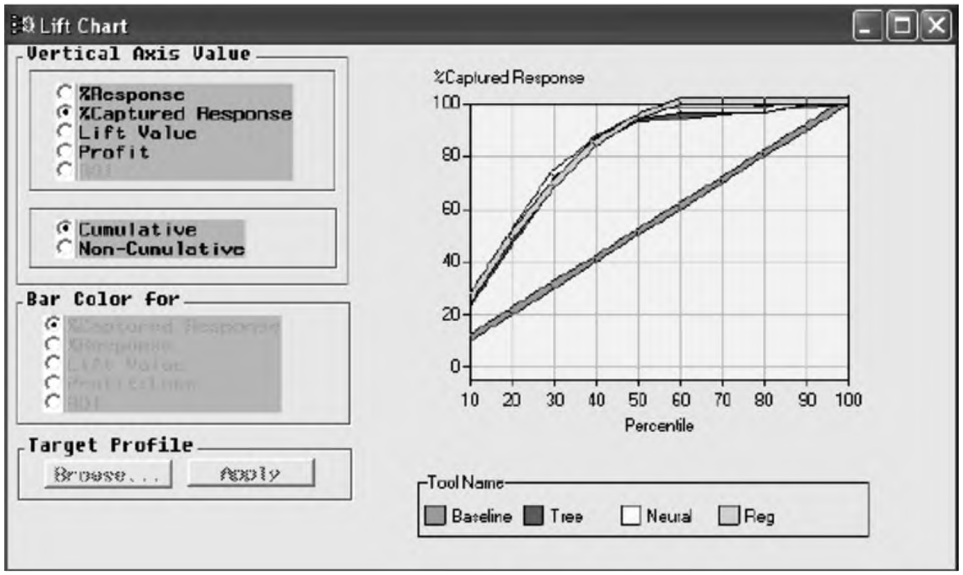


Figura 12-35

- ***Lift Value***: muestra la ganancia que obtenemos al aplicar el modelo igual que si tomáramos la decisión al azar (Figura 12-36). Este valor está especialmente indicado cuando se especifica una matriz de beneficios o pérdidas.

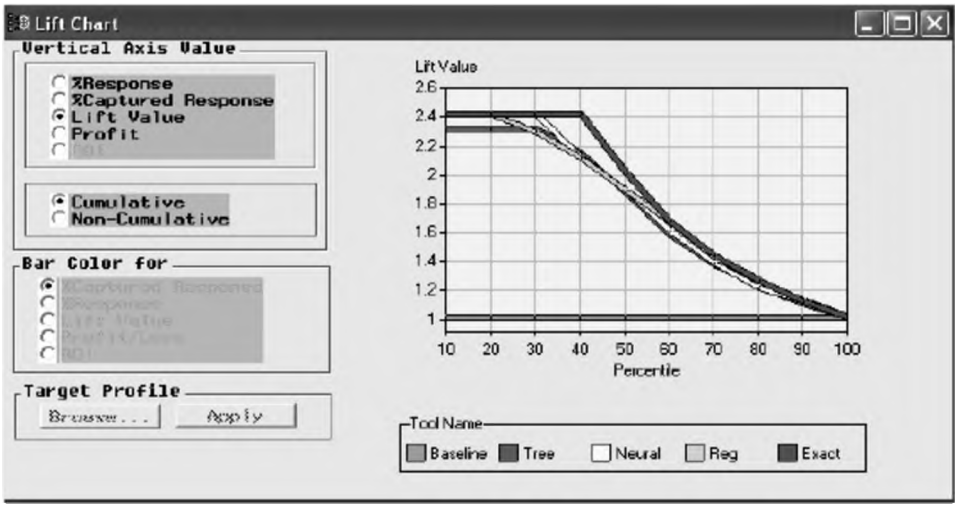


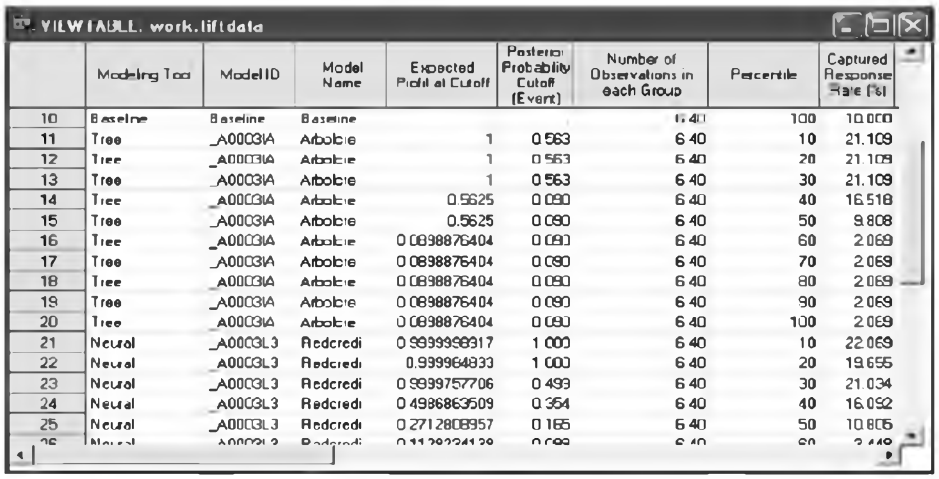
Figura 12-36

- ***Profit:*** el gráfico Profit muestra el beneficio o pérdida acumulado (o no acumulado) para cada percentil.

- ***ROI (Return on Investment)***: este gráfico muestra el rendimiento de la inversión acumulado o no acumulado para cada decila. La tasa de rendimiento se define como el ratio de los beneficios y los costes expresados en forma porcentual. Por ejemplo si una campaña de publicidad cuesta 20000 euros y esperamos con ella obtener unos beneficios de 45000 euros en las dos primeras decilas entonces el valor acumulado ROI será (45000-20000)/20000 cuyo resultado multiplicaremos por 100 → 125%. Para crear este gráfico debemos definir un coste constante o variable en la matriz de costes del *Target Profile*.

Vista de los datos relativos a los gráficos de ganancias

Cada vez que un gráfico es creado *Enterprise Miner* genera un fichero de datos temporales con el nombre WORK.LIFTDATA. Este gráfico contiene las estadísticas por decilas de cada modelo. Para ver este fichero seleccionaremos *View Lift Data* en la barra de herramientas (Figura 12-37).



	Modeling Tool	Model ID	Model Name	Expected Profit at Cutoff	Posterior Probability Cutoff (Event)	Number of Observations in each Group	Percentile	Captured Response Rate (%)
10	Baseline	Baseline	Baseline			6.40	100	10.000
11	Tree	_A00C3IA	Arbolcie	1	0.563	6.40	10	21.109
12	Tree	_A00C3IA	Arbolcie	1	0.563	6.40	20	21.109
13	Tree	_A00C3IA	Arbolcie	1	0.563	6.40	30	21.109
14	Tree	_A00C3IA	Arbolcie	0.5625	0.090	6.40	40	16.518
15	Tree	_A00C3IA	Arbolcie	0.5625	0.090	6.40	50	9.808
16	Tree	_A00C3IA	Arbolcie	0.0898876404	0.090	6.40	60	2.069
17	Tree	_A00C3IA	Arbolcie	0.0898876404	0.090	6.40	70	2.069
18	Tree	_A00C3IA	Arbolcie	0.0898876404	0.090	6.40	80	2.069
19	Tree	_A00C3IA	Arbolcie	0.0898876404	0.090	6.40	90	2.069
20	Tree	_A00C3IA	Arbolcie	0.0898876404	0.090	6.40	100	2.069
21	Neural	_A00C3L3	Redcredi	0.9999998917	1.000	6.40	10	22.069
22	Neural	_A00C3L3	Redcredi	0.999964833	1.000	6.40	20	19.655
23	Neural	_A00C3L3	Redcredi	0.9999757706	0.499	6.40	30	21.034
24	Neural	_A00C3L3	Redcredi	0.4986863509	0.354	6.40	40	16.092
25	Neural	_A00C3L3	Redcredi	0.2712808957	0.165	6.40	50	10.805

Figura 12-37

El fichero puede ser exportado de la forma descrita con anterioridad en *File → Export*.

DIAGNOSTIC (Gráfico de clasificación)

Este gráfico permite comprobar el ajuste de un solo modelo a la vez. Por defecto el punto de corte se establece en 0,5. Si queremos variar este punto deberemos trazar un gráfico de tipo *threshold-based* cuyas opciones veremos posteriormente (Figura 12-38).

©Alfaomega-RC Libros

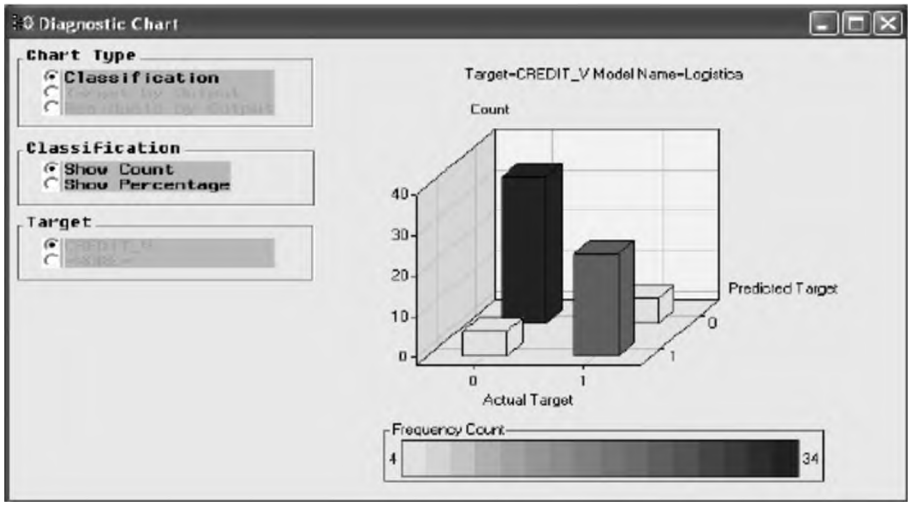


Figura 12-38

Si el modelo es útil las barras relativas a los casos acertados serán más altas que las barras de los errores. En el caso de variables continuas podremos trazar *Target by Output* (valores reales *vs* valores predichos) mostrado en la Figura 12-39 y *Residuals by output* (Errores vs valores reales).

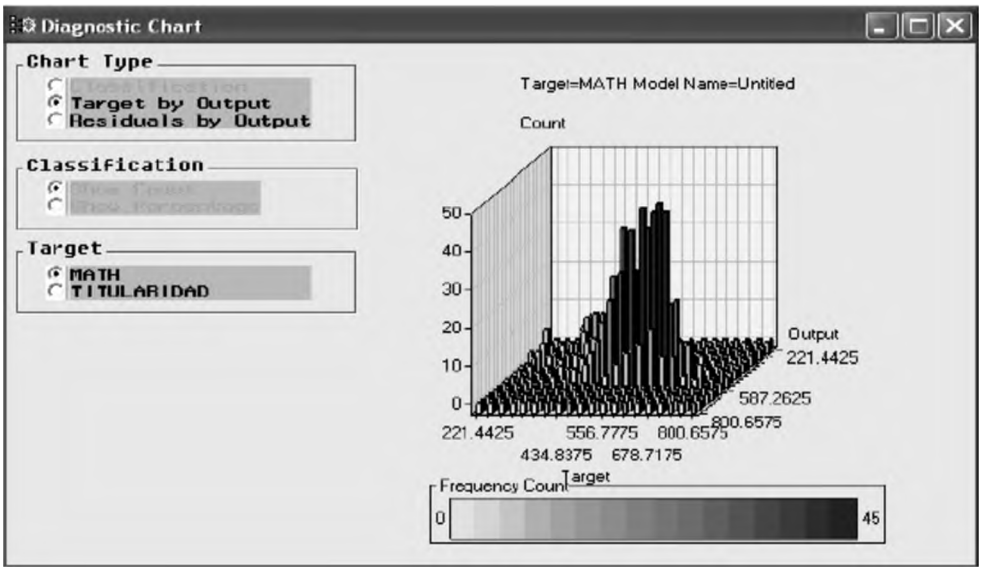


Figura 12-39

THRESHOLD-BASED (Gráfico de clasificación basado en el umbral)

Para acceder a esta opción en el menú principal seleccionaremos *Tools* → *Threshold-based*. Esta opción permite representar tres tipos de gráficos.

Standard Threshold-based

Muestra el gráfico de la matriz de confusión según el punto de corte (Figura 12-40).

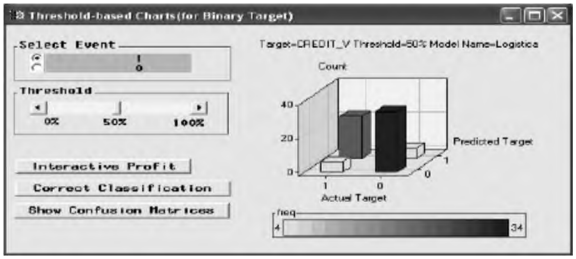


Figura 12-40

El gráfico permite modificar de forma interactiva la categoría de referencia (*Event*) y el umbral (*Threshold*).

Interactive Profit

Para activar esta opción seleccionaremos la opción del mismo nombre del gráfico anterior. Este gráfico permite medir el impacto que la matriz de pérdidas y ganancias tiene en el rendimiento promedio a lo largo de distintos puntos de corte (Figura 12-41). La matriz de beneficios se muestra a la izquierda mientras que el rendimiento según el umbral se muestra a la derecha. La matriz de beneficios permite definir un beneficio o pérdida para cada clasificación.

Correct Classification (Clasificaciones correctas)

Muestra el ajuste para cada nivel de respuesta y la combinación de ambas (curva amarilla) a medida que varía el punto de corte (Figura 12-42). Estos gráficos solo representan el ajuste del modelo sin hacer referencia a la matriz de beneficio. El valor de los puntos de la curva puede ser visualizado con la herramienta *View Info*.

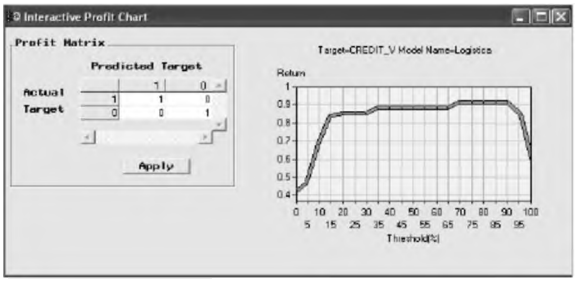


Figura 12-41

©Alfaomega-RC Libros

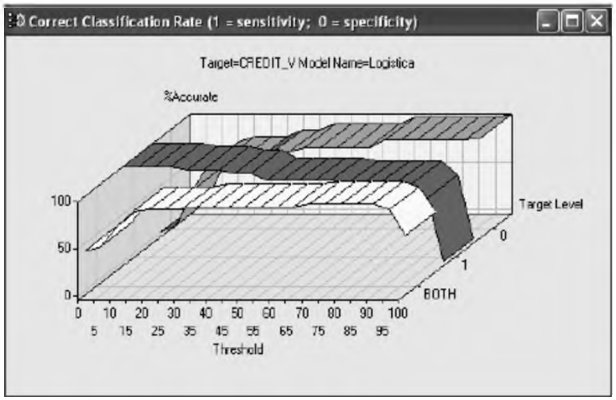
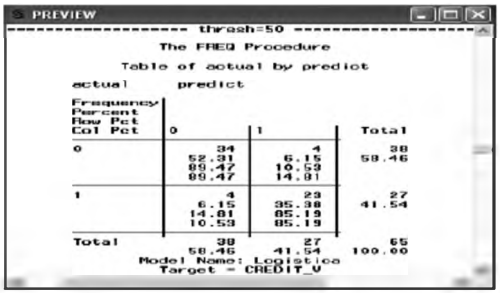


Figura 12-42

Confusion Matrix (Matriz de Confusión)

Muestra el porcentaje y la naturaleza del error (Figura 12-43). La matriz de confusión varía con los puntos de corte y la distinta cuantía de los errores y su variación determinan las coordenadas de las curvas ROC (*Receiver Operating Characteristics*). La opción *Show Confusion Matrix* muestra las matrices de confusión variando el umbral de 5 en 5 por ciento hasta un total de 21 matrices.



PREVIEW

thresh=50

The FREQ Procedure

Table of actual by predict

actual / predict (Frequency, Percent, Row Pct, Col Pct)	0	1	Total
0	34 52.31 89.47 89.47	4 6.15 10.53 14.81	38 58.46
1	4 6.15 14.81 10.53	23 35.38 85.19 85.19	27 41.54
Total	38 58.46	27 41.54	65 100.00

Model Name: Logistica
Target = CREDIT_V

Figura 12-43

La tabla anterior muestra cómo para un punto de corte de 0,5 (Thresh=50) el modelo de regresión logística tiene 8 errores en la matriz de validación que consta de 65 casos.

Curvas ROC

Este tipo de gráfico es útil para comparar el comportamiento global de un modelo y solo es posible su representación con variables respuesta binarias. Este gráfico enfrenta dos variables: la sensibilidad (sensitivity) y 1 – especificidad (1 – specificity).

- ***Sensitivity***: es una medida de la capacidad de acierto de un evento y se define como el número de categorías positivas (Valor 1, YES, POSITIVE o categoría de referencia) bien predichas dividido por el total de categorías positivas.

- **Specificity**: es una medida de la capacidad de acierto del evento complementario al anterior. Se define como el número de categorías falsas (valor 0, NO, NEGATIVE o categoría complementaria a la de referencia) bien predichas (el modelo también dice que son falsas) dividido por el total de categorías falsas. 1 – Specificity es simplemente el número de falsos positivos o número de observaciones con categoría falsa que el modelo incorrectamente predice como verdaderas para un punto de corte determinado dividido por el número de casos falsos.

La curva muestra el trade-off entre ambas magnitudes (Figura 12-44). Idealmente el objetivo es tener altas ambas para predecir correctamente ambas categorías. Un punto de corte más bajo aumenta el número de positivos falsos mientras que un punto de corte más alto aumenta el número de negativos falsos.

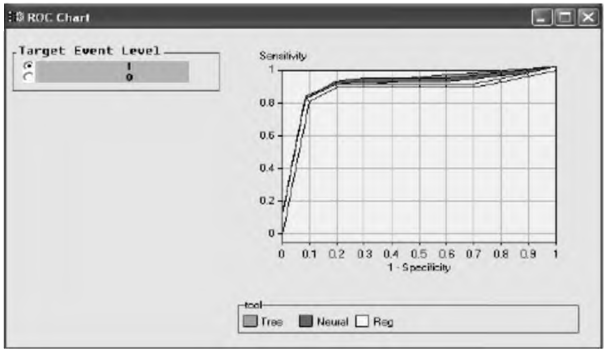


Figura 12-44

Scatter Plot (Gráfico de dispersión)

Con variables objetivo continuas o de intervalo podemos crear gráficos de dispersión. Para ello seleccionaremos *Tools → Predicted Plot*. El gráfico permite variar las variables que queremos posicionar tanto en ordenadas como en abscisas (Figura 12-45).

©Alfaomega-RC Libros

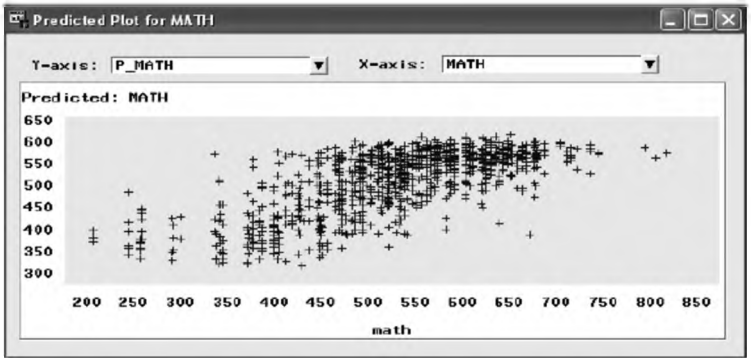


Figura 12-45

EJEMPLO DE COMBINACIÓN DE MODELOS DE ÁRBOLES CON OTROS MODELOS

En esta ocasión ilustraremos cómo combinar las respuestas de un modelo de redes neuronales y de un modelo de árboles de decisión a través del nodo *ensemble*. Para ello utilizaremos los datos del archivo CREDITOS. Una vez importado el fichero, asignaremos a la variable Cliente el papel ID en el modelo mientras que la variable CREDIT_V será la variable objetivo (*Target*). El resto de variables serán inputs. A continuación uniremos el nodo de datos con el de partición y asignaremos de manera aleatoria un 80% de los datos a entrenamiento, un 10% a validación y un 10% a test. Una vez realizada esta tarea conectaremos el nodo de partición de datos con un nodo de redes neuronales (Neural Network) y con un nodo de árboles de decisión (Tree) según se indica en la Figura 12-46.

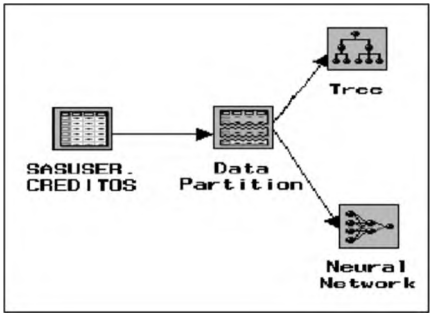


Figura 12-46

Una vez conectados, ejecutaremos, dejando las opciones por defecto, ambos nodos. A continuación uniremos ambos modelos al nodo de combinación de modelos (Figura 12-47).

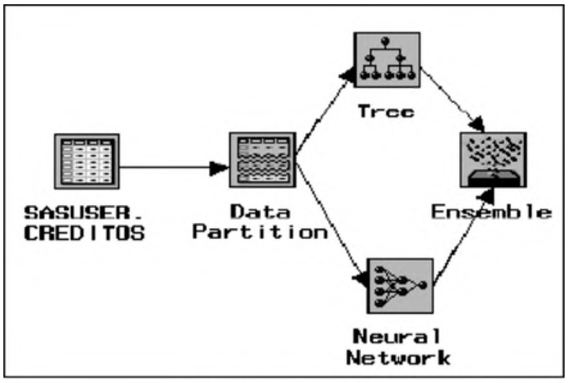


Figura 12-47

En la ventana *Setting* la mejor opción es que el mismo nodo detecte la opción más adecuada para combinar los resultados (*Automatic*) según se muestra en la Figura 12-48.

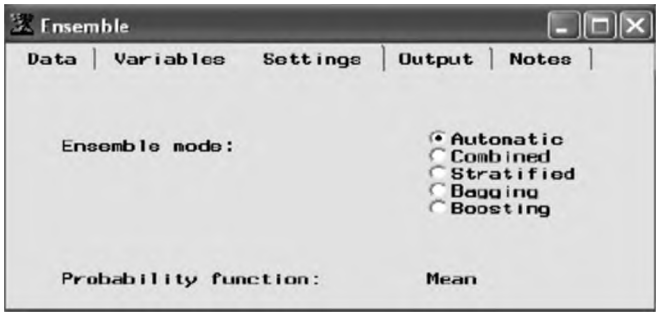


Figura 12-48

En la ventana output indicamos que procese los resultados para los datos de entrenamiento, validación y test (Figura 12-49).

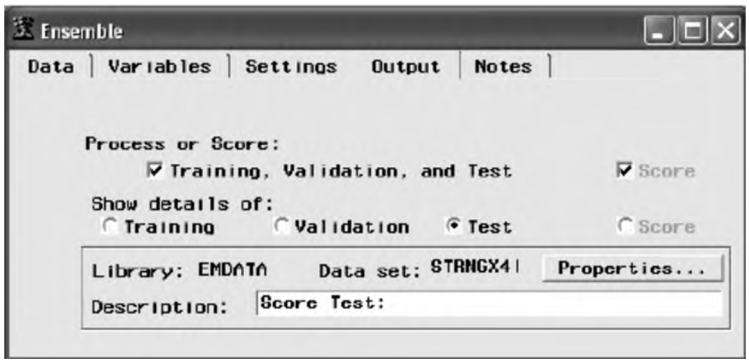


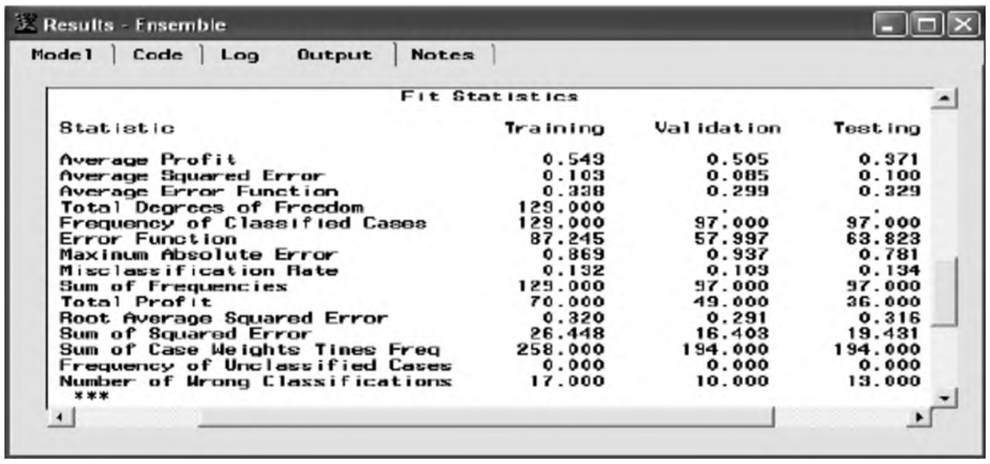
Figura 12-49

Antes de ejecutar el nodo SAS Enterprise Miner nos pedirá que asignemos un nombre al modelo combinado (Figura 12-50).

©Alfaomega-RC Libros

Figura 12-50

Una vez ejecutado el nodo, el visor de resultados mostrará en la pestaña Output las estadísticas descriptivas resultantes de la combinación (Figura 12-51).



Fit Statistics

Statistic	Training	Validation	Testing
Average Profit	0.543	0.505	0.371
Average Squared Error	0.103	0.085	0.100
Average Error Function	0.338	0.299	0.329
Total Degrees of Freedom	129.000	.	.
Frequency of Classified Cases	129.000	97.000	97.000
Error Function	87.245	57.997	63.823
Maximum Absolute Error	0.869	0.937	0.781
Misclassification Rate	0.132	0.103	0.134
Sum of Frequencies	129.000	97.000	97.000
Total Profit	70.000	49.000	36.000
Root Average Squared Error	0.320	0.291	0.316
Sum of Squared Error	26.448	16.403	19.431
Sum of Case Weights Times Freq	258.000	194.000	194.000
Frequency of Unclassified Cases	0.000	0.000	0.000
Number of Wrong Classifications	17.000	10.000	13.000

***

Figura 12-51

Para valorar estos resultados podemos conectar un nodo de valoración *Assessment* a los tres modelos (Figura 12-52).

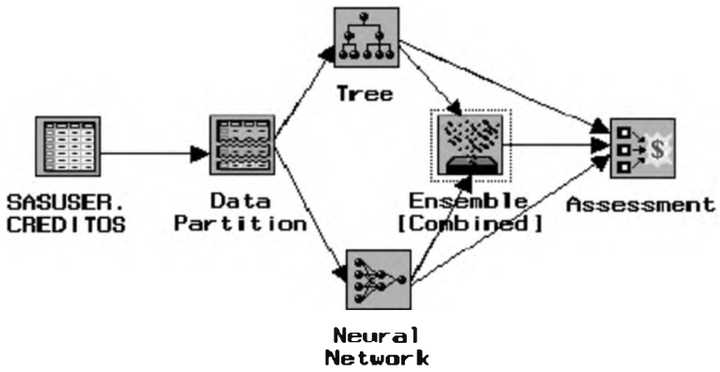


Figura 12-52

Si abrimos el nodo *Assessment* los resultados muestran que la combinación ofrece unos resultados que pueden llegar a mejorar ligeramente los resultados de los dos modelos originarios. Este resultado es posible cuando cada modelo original funciona mejor o peor que el otro dependiendo del tramo de la función que estemos evaluando.

A

Almacén de datos *2, 3, 4, 5, 6, 7, 8, 9, 10, 13, 14*
Almacén de datos y las herramientas OLAP *13*
Almacenes de datos *4*
Análisis cluster *183, 184, 185, 187, 193, 199*
Análisis cluster como técnica de perfilado y segmentación *183*
Análisis cluster con redes neuronales *204*
Análisis cluster con redes neuronales *236*
Análisis cluster con sas enterprise miner *229*
Análisis cluster no jerárquico *186*
Análisis de conglomerados. *182, 183, 184, 185, 194, 200, 204, 209, 229*
Análisis de la covarianza múltiple *127*
Análisis de la covarianza simple *126*
Análisis de la regresión múltiple *124, 128*
Análisis de la varianza múltiple *127*
Análisis de la varianza simple *126*
Análisis discriminante *125, 128, 155, 157, 182, 184, 211, 214, 376*
Analysis Services *15, 91, 92, 93, 94, 95, 96, 100, 101, 112, 117, 118, 119, 120, 121, 122*
Árboles cart *213*
Árboles chaid *212*
Árboles de clasificación y regresión *214*
Árboles de decisión *4, 122, 128, 129, 131, 175, 182, 208, 209, 214, 215, 218, 221, 224, 244, 252, 257, 261, 262, 268, 375, 376, 385, 411*
Árboles quest *213*

B

Base de datos transaccional *4*
Business intelligence (BI) en SharePoint *91*
Business intelligence en Microsoft BI SQL Server .. *91*
Búsqueda rápida en bing *86*

C

Características de Power BI *26, 27*

C

Chi-square automatic interaction detector *212*
Classification and regression trees *213*
Cluster jerárquico *199*
Clustering mediante redes neuronales *189*
Clusters jerárquicos, secuenciales, aglomerativos y exclusivos *185*
Consultas avanzadas *3*
Crear jerarquías *59*
Crear relaciones *58*
Criterios para minimizar los costes *210*
CUBE y rollup en SQL *15*

D

Data warehouse *4*
Dendograma en el análisis cluster *185*
Detector automático de interacciones *212*

E

EIS (executive information system) *2*
ETL (extraer, transformar y cargar) *3*
Excel Power Map *89*
Excel services en SharePoint server 2013 ..... *93*
Exportar a powerpoint desde Power View *99*

F

Finalidad de los sistemas de información *1*
Funciones de cubo en Excel *22*

G

Gráfico de burbujas *107*
Gráficos circulares *106*
Gráficos de barras *108*
Gráficos de columnas *109*
Gráficos de dispersión y de burbujas *107*
Gráficos de líneas *108, 109*
Gráficos de líneas, barras y columnas *108*
Gráficos y otras visualizaciones en Power View .. *104*
Granja de servidores de SharePoint *92, 93, 95*

H

Herramientas OLAP *2, 10, 11, 13*
Herramientas y técnicas ELT *3*

I

Informe de Power View *64, 76, 83, 98*
Informes con Power BI *63*
Informes con segmentaciones de datos *79*
Informes de comparaciones *82*
Informes de Power View *100*
Informes de reporting services *113*
Informes de SQL server reporting services *113*
Informes detallados *75*
Informes OLAP con CUBE *15*
Informes OLAP con CUBE y ROLLUP en SQL *15*
Integration services *91, 112, 116, 117, 118, 122*
Interfaces multidimensionales *2*

M

Mapas *109*
Mapas de Power View *109*
Método Cart *213*
Método Chaid *212*
Método de división *211*
Método Quest *214*
Métodos de tipo discriminante *211*
Microsoft BI SQL Server *91*
Microsoft Power BI para Office 365 ..... *25*
Microsoft SharePoint *91, 93, 96*
Modelado predictivo *123*
Modelización automática *367*
Modelo logit *126*
Modelo probit *126*
Modelos de regresión con variables ficticias *124*
Modelos tabulares son bases de datos "en memoria" de analysis services *120*
Mosaicos en Power View *111*

O

On-line analyitical processing *2*
On-line transactional processing *4*
Operaciones de OLAP típicas *12*

P

Permisos para Power View *98*
Plataforma Microsoft BI *91*
Poda *211*
Ponderaciones de clase *210*
Power BI *25*
Power Map *27, 86, 89*
Power Pivot27, 41, 44, 45, 46, 54, 55, 56, 57, 58, 59, 60, 61, 65, 66, 68, 82, 98
Power query.... 26, 27, 28, 30, 32, 33, 34, 35, 37, 40, 42, 46, 50, 53, 54, 61, 64
Power View27, 61, 63, 64, 65, 66, 68, 69, 71, 75, 76, 78, 79, 82, 83, 91, 93, 96, 97, 98, 99, 100, 101, 102, 104, 106, 108, 109, 111, 120
Power View para excel 2013 ..... *99*
Power View para SharePoint *96*
Power View para SharePoint server.. 96, 97, 99, 100
Powerpivot para excel *91, 93, 95, 101, 119*
Powerpivot para SharePoint..91, 92, 93, 94, 95, 119
Probabilidades a priori *210*
Problema del sobreajuste *211*

R

Redes neuronales de Kohonen *204*
Reglas de asociación *267, 269, 276*
Regresión logística *126, 141, 146*
Reporting Services 91, 92, 93, 96, 99, 112, 113, 118, 119, 121
Rollup en SQL *15*

S

Segmentación ad-hoc *182*
Segmentación post hoc *182, 183*
Segmentación post-hoc *182*
Segmentación y perfilado con árboles de decisión *208*
Segmentaciones de datos *79*
SharePoint 2013 ..... *92, 93, 94, 104*
Sistemas de información para la gestión *1*
Sistemas de información para la toma de decisiones *1*
Sistemas de informes *3*
Sistemas HOLAP *14*
Sistemas MOLAP *13*

©Alfaomega-RC Libros

Sistemas ROLAP *13, 14*
SQL server integration services *116*

T

Técnicas de modelado para la segmentacion..... *181*
Técnicas de modelado predictivo..... *123*
Técnicas OLAP *3*
Transact SQL..... *15*

V

Visualización con Power View *63*
Visualizaciones en Power View.....*104*

W

Whit ROLLUP *20*
With CUBE..... *16, 17, 19, 20, 21, 22*